Estadística descriptiva univariante

Modelos estadísticos para la descripción de datos univariantes

Alicia Vila y Ángel A. Juan

PID_00161058



Índice

Introducción 5

Objetivos 6

1. Introducción a la Estadística 7

2. Descripción de datos mediante tablas y gráficos 11

3. Descripción de datos mediante estadísticos 18

4. El concepto de probabilidad 25

5. Distribuciones de probabilidad discretas 28

6. Distribuciones de probabilidad continuas 35

Resumen 45

Ejercicios de autoevaluación 47

Solucionario 49

Introducción

Las sociedades modernas son ricas en datos: la prensa escrita, la televisión y la radio, Internet y las intranets de las organizaciones ofrecen cantidades inmensas de datos que pueden ser procesados y analizados. Esto convierte a la estadística en una ciencia interesante y útil puesto que proporciona estrategias y herramientas que permiten obtener información a partir de dichos datos. Además, gracias a la evolución de la tecnología (ordenadores y software estadístico) hoy en día es posible automatizar gran parte de los cálculos matemáticos asociados al uso de técnicas estadísticas, lo que permite extender su uso a un gran rango de profesionales en ámbitos tan diversos como la biología, las ciencias empresariales, la sociología o las ciencias de la información.

La práctica de la estadística requiere aprender a obtener y explorar los datos –tanto numéricamente como mediante gráficos–, a pensar sobre el contexto de los datos y el diseño del estudio que los ha generado, a considerar la posible influencia de observaciones anómalas en los resultados obtenidos, a discutir la legitimidad de los supuestos requeridos por cada técnica y, finalmente, a validar la fiabilidad de las conclusiones derivadas del análisis. La estadística requiere tanto de conocimientos sobre los conceptos y técnicas empleados como de la suficiente capacidad crítica que permita evaluar la conveniencia de usar unas u otras técnicas según el tipo de datos disponible y el tipo de información que se desea obtener.

En este módulo inicial de la asignatura, se examinan los datos procedentes de una única variable: en primer lugar se explica cómo organizar y resumir dichos datos, tanto numérica como gráficamente (estadística descriptiva); en segundo lugar, se introducen los conceptos básicos asociados con la idea de probabilidad; finalmente, se presentan algunos modelos matemáticos que permiten analizar el comportamiento de algunas variables.

Objetivos

Los objetivos académicos que se plantean en este módulo son los siguientes:

1. Entender la importancia de la estadística en la sociedad moderna.

2. Aprender a organizar y resumir un conjunto de datos procedentes de una variable mediante gráficos, tablas de frecuencias y estadísticos descriptivos.

3. Comprender el concepto de probabilidad de un suceso y descubrir sus principales propiedades y aplicaciones.

4. Conocer las principales distribuciones estadísticas que se usan para modelar el comportamiento de variables discretas y continuas.

5. Saber calcular probabilidades asociadas a cada una de las distribuciones introducidas.

6. Aprender a usar software estadístico o de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y técnicas estadísticas.

1. Introducción a la Estadística

La Estadística es la ciencia que se ocupa de obtener datos y procesarlos para transformarlos en información. Es, por tanto, un lenguaje universal ampliamente utilizado en las ciencias sociales, en las ciencias experimentales, en las ciencias de la salud y en las ingenierías. Las Tecnologías de la Información y la Comunicación (TIC) han incrementado notablemente la producción, diseminación y tratamiento de la información estadística. En particular, Internet es una fuente inagotable de datos que pueden ofrecer información y, a partir de ella, conocimiento. Por otra parte, la constante evolución de los ordenadores personales y de los **programas informáticos de estadística** y análisis de datos posibilita y facilita el análisis de grandes cantidades de datos mediante el uso de técnicas estadísticas y de minería de datos. En la Sociedad de la Información se hace pues imprescindible disponer de un cierto conocimiento estadístico incluso para poder comprender e interpretar correctamente los indicadores económicos (IPC, inflación, tasa de desempleo, Euribor, etc.), los indicadores bibliométricos (factor de impacto de una revista, cuartil en el que se sitúa, vida media de las citas recibidas, etc.) o los indicadores sociales (esperanza de vida, índice de alfabetización, índice de pobreza, indicador social de desarrollo sostenible, etc.) a los que frecuentemente se hace referencia en los medios de comunicación.

Nota

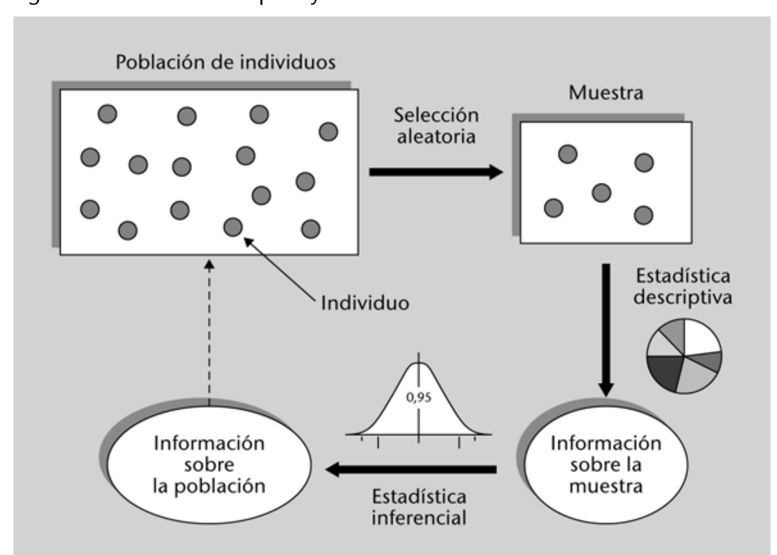
Las agencias gubernamentales, como el Instituto Nacional de Estadística (INE) o el Eurostat proporcionan datos sobre casi cualquier ámbito socioeconómico.

Software estadístico

En la actualidad existen excelentes **programas informáticos** para el análisis estadístico de datos. Algunos ejemplos son: MINITAB, SPSS, MS Excel, SAS, R, S-Plus, Statgraphics o Statistica.

El campo de la Estadística se puede dividir en dos grandes áreas: la estadística descriptiva y la estadística inferencial (figura 1).

Figura 1. Estadística descriptiva y estadística inferencial



La estadística descriptiva se ocupa de la obtención, presentación y descripción de datos procedentes de una muestra o subconjunto de una población de individuos. Por su parte, la estadística inferencial usa los resultados obtenidos

mediante la aplicación de las técnicas descriptivas a una muestra para inferir información sobre el total de la población a la que pertenece dicha muestra.

Algunos términos básicos

A lo largo de este material se usarán abundantes términos estadísticos, muchos de ellos bastante conocidos. A continuación se presentan y revisan algunos de estos términos básicos que conviene entender bien:

- **Población:** colección o conjunto de elementos (individuos, objetos o sucesos) cuyas propiedades se desean analizar. Ejemplos: (a) los estudiantes universitarios de un país; (b) el conjunto de periódicos en Internet; (c) el conjunto de revistas indexadas en el Science Citation Index (SCI), etc.

- **Muestra:** cualquier subconjunto de elementos de la población. Ejemplos: (a) los estudiantes de una determinada universidad; (b) los periódicos en línea centrados en aspectos económicos; (c) las revistas indexadas en el SCI de una determinada editorial, etc.

- **Muestra aleatoria:** muestra cuyos elementos han sido escogidos de forma aleatoria. Ejemplos: (a) un subconjunto de doscientos estudiantes escogidos al azar (mediante el uso de números aleatorios) de entre todos los matriculados en universidades de un país; (b) un subconjunto de cincuenta periódicos en línea escogidos al azar; (c) un subconjunto de quince revistas indexadas en el SCI escogidas al azar, etc.

- **Marco del muestreo:** lista que contiene aquellos elementos de la población candidatos a ser seleccionados en la fase de muestreo. No necesariamente coincidirá con toda la población de interés, ya que en ocasiones no será posible identificar a todos los elementos de la población. Ejemplos: (a) lista de todos los estudiantes matriculados en universidades de un país en un semestre concreto; (b) relación de periódicos en línea disponibles en un momento dado; (c) lista de todas las revistas indexadas en el SCI en un año específico, etc.

- **Variable aleatoria:** característica de interés asociada a cada uno de los elementos de la población o muestra considerada. Ejemplos: (a) la edad de cada estudiante; (b) el número de visitas diarias que recibe cada periódico en línea; (c) el factor de impacto de cada revista, etc.

- **Datos u observaciones:** conjunto de valores obtenidos para la variable de interés en cada uno de los elementos de la muestra. Ejemplos: (a) las edades registradas son {25, 23, 19, 28…}; (b) las visitas diarias registradas son {1326, 1792, 578, 982…}; (c) los factores de impacto registrados son {2,3; 1,7; 8,2…}.

- **Experimento:** estudio en la que el investigador controla o modifica expresamente las condiciones del mismo con la finalidad de analizar los distin-

tos patrones de respuesta en las observaciones. Ejemplos: (a) estudiar cómo varían las calificaciones de un grupo de estudiantes según dispongan o no de ordenadores con acceso a Internet en las aulas; (b) estudiar cómo varía el número de visitas a un periódico en línea según se opte o no por incluir noticias sensacionalistas en su portada; (c) estudiar cómo varía el factor de impacto de un grupo de revistas según éstas se incluyan o no en una base de datos de reconocido prestigio, etc.

- **Inspección o encuesta:** estudio en el que el investigador no pretende modificar las condiciones de la muestra con respecto a la variable de interés sino simplemente obtener los datos correspondientes a unas condiciones estándar. Ejemplos: (a) registrar las calificaciones de los estudiantes de un máster determinado; (b) realizar una encuesta a los lectores de un periódico en línea; (c) obtener el factor de impacto asociado a cada una de las revistas de una muestra, etc.

- **Parámetro:** valor numérico que sintetiza alguna propiedad determinada de la población. Los parámetros se asocian a toda la población y suelen representarse con letras del alfabeto griego como μ (mu), σ (sigma), etc. Ejemplos: (a) la edad media de todos los estudiantes universitarios de un país; (b) el número máximo de visitas diarias recibido por algún periódico en línea; (c) el rango o diferencia entre el mayor y el menor factor de impacto del conjunto de revistas indexadas en el SCI, etc.

- **Estadístico:** valor numérico que sintetiza alguna propiedad determinada de una muestra. Los estadísticos se asocian a una muestra y se suelen representar por letras del alfabeto latino como $\overline{x}$, s, etc. Ejemplos: (a) la edad media de los estudiantes de una muestra aleatoria; (b) el número máximo de visitas diarias recibidas por algún periódico deportivo en línea; (c) el rango o diferencia entre el mayor y el menor factor de impacto de las revistas de una editorial, etc.

- **Variable cualitativa o categórica:** variable que categoriza o describe cualitativamente un elemento de la población. Suele ser de tipo alfanumérico, pero incluso en el caso en que sea numérica no tiene sentido usarla en operaciones aritméticas. Ejemplos: (a) el teléfono o el correo electrónico de un estudiante; (b) la dirección IP de un periódico en línea; (c) el ISSN de una revista, etc.

- **Variable cuantitativa o numérica:** variable que cuantifica alguna propiedad de un elemento de la población. Es posible realizar operaciones aritméticas con ella. Ejemplos: (a) el importe de la beca que recibe un estudiante; (b) los ingresos que genera un periódico en línea; (c) el número de revistas publicadas por una editorial, etc.

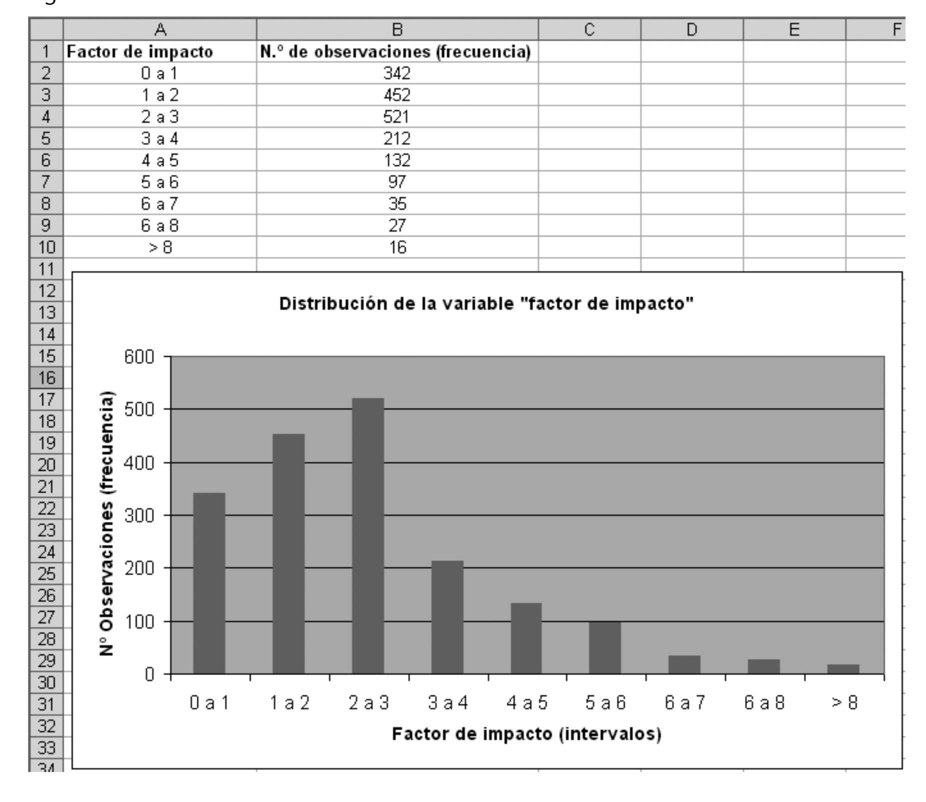
- **Variable cuantitativa discreta:** variable cuantitativa que puede tomar un número finito o contable de valores distintos. Ejemplos: (a) edad de un es-

tudiante; (b) número de enlaces a otras fuentes de información que ofrece un periódico en línea; (c) calificación que obtiene una revista en una escala entera de 1 a 5, etc.

- **Variable cuantitativa continua:** variable cuantitativa que puede tomar un número infinito (no contable) de valores distintos. Ejemplos: (a) altura o peso de un estudiante; (b) tiempo que transcurre entre la publicación de una encuesta en línea y el instante en que ya la han completado un centenar de internautas; (c) factor de impacto (sin redondear) de una revista, etc.

- **Distribución de una variable:** en sentido amplio, una distribución es una tabla, gráfico o función matemática que explica cómo se comportan o distribuyen los valores de una variable, es decir, qué valores toma la variable así como la frecuencia de aparición de cada uno de ellos. Ejemplo: dada una muestra aleatoria de revistas, la distribución de la variable "factor de impacto de una revista" puede representarse mediante una tabla de frecuencias o mediante una gráfica como se aprecia en la figura 2. Se observa que trescientas cuarenta y dos de las revistas consideradas tienen un factor de impacto entre 0 y 1, cuatrocientas cincuenta y dos de las revistas tienen un factor de impacto entre 1 y 2, etc.

Figura 2. Distribución de una variable aleatoria



	A	B	C	D	E	F
1	**Factor de impacto**	**N.° de observaciones (frecuencia)**				
2	0 a 1	342				
3	1 a 2	452				
4	2 a 3	521				
5	3 a 4	212				
6	4 a 5	132				
7	5 a 6	97				
8	6 a 7	35				
9	6 a 8	27				
10	> 8	16				

2. Descripción de datos mediante tablas y gráficos

Cuando se dispone de un conjunto de observaciones procedentes de una muestra conviene hacer un primer análisis exploratorio de éstas mediante gráficos y tablas que ayuden a interpretar los datos y a extraer información de los mismos. Existen diferentes tipos de gráficos que pueden usarse en esta fase exploratoria y el uso de unos u otros dependerá en gran medida del tipo de datos de los que se disponga (cualitativos o cuantitativos), así como de la información que se desee visualizar. En este apartado se presentaran algunos de los gráficos y tablas más habituales para la descripción de **datos univariantes**.

Datos univariantes

Los datos univariantes son los que provienen de una única variable. En algunos casos, los datos pueden proceder de dos o más variables y, entonces, se usa la expresión bivariante (si se trata de dos variables) o multivariante (si se consideran más de dos).

Gráficos y tablas para datos cualitativos o categóricos

Si se dispone de datos cualitativos o categóricos, pueden sintetizarse mediante una tabla que recoja, para cada categoría: el número de veces que aparece (frecuencia absoluta), el porcentaje de apariciones sobre el total de observaciones (frecuencia relativa), así como los acumulados de ambos valores. La tabla 1 muestra esta información para la variable "número de *hotspots* (conexiones *wi-fi*) identificados en cada comunidad autónoma".

Tabla 1. Ejemplo de tabla de frecuencias para una variable categórica

Comunidad autónoma	***Hotspots* por comunidad autónoma**			
	Frecuencia	**Frecuencia acumulada**	**Frecuencia relativa**	**Frec. rel. acumulada**
Andalucía	885	885	11,9%	11,9%
Aragón	177	1.062	2,4%	14,2%
Asturias	148	1.210	2,0%	16,2%
Cantabria	164	1.374	2,2%	18,4%
Castilla-La Mancha	144	1.518	1,9%	20,3%
Castilla y León	302	1.820	4,0%	24,4%
Cataluña	1.391	3.211	18,6%	43,0%
C. Valenciana	622	3.833	8,3%	51,3%
Extremadura	137	3.970	1,8%	53,2%
Galicia	516	4.486	6,9%	60,1%
I. Baleares	183	4.669	2,5%	62,5%
I. Canarias	151	4.820	2,0%	64,6%
La Rioja	126	4.946	1,7%	66,3%
Madrid	1.776	6.722	23,8%	90,0%
Murcia	160	6.882	2,1%	92,2%
Navarra	153	7.035	2,0%	94,2%
País Vasco	430	7.465	5,8%	100,0%
Totales	**7.465**		**100,0%**	

Nota

Observad que la **frecuencia acumulada** se obtiene sólo con ir acumulando frecuencias anteriores.

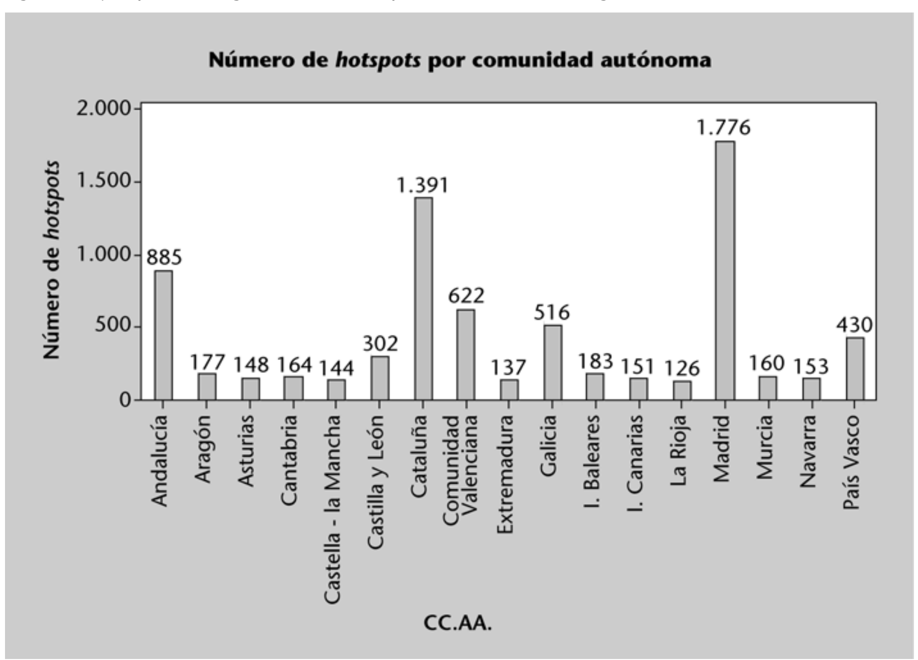
Además de mediante una tabla de frecuencias, suele ser habitual representar datos categóricos mediante el uso de gráficos circulares (figura 3) o bien mediante diagramas de barras (figura 4).

Figura 3. Ejemplo de gráfico circular para una variable categórica

Porcentaje de *hotspots* por comunidades autónomas

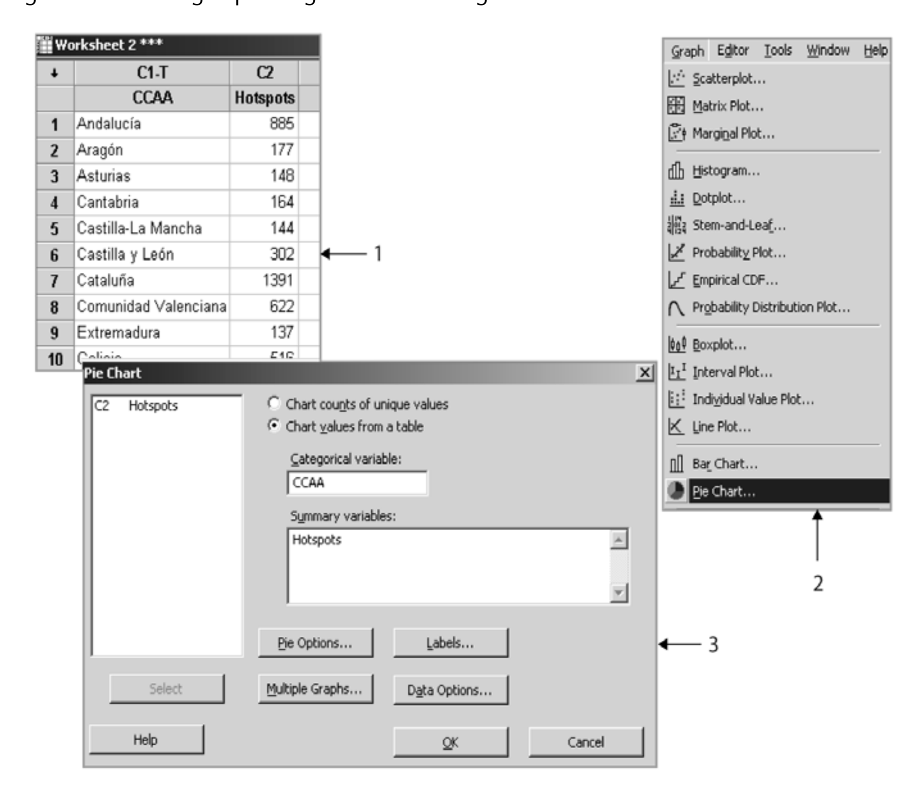
Navarra 2,0%
País Vasco 5,8%
Andalucía 11,9%
Aragón 2,4%
Asturias 2,0%
Cantabria 2,2%
Castilla – La Mancha 1,9%
Castilla y León 4,0%
Cataluña 18,6%
Comunidad Valenciana 8,3%
Extremadura 1,8%
Galicia 6,9%
I. Baleares 2,5%
I. Canarias 2,0%
La Rioja 1,7%
Madrid 23,8%
Murcia 2,1%

Figura 4. Ejemplo de diagrama de barras para una variable categórica



Este tipo de gráficos pueden crearse fácilmente con cualquier programa estadístico o de análisis de datos (p. ej.: Minitab, MS Excel, SPSS, etc.). La figura 5 muestra los pasos básicos para generar un gráfico circular (*pie chart*) con Minitab. La generación de un diagrama de barras (*bar chart*) se consigue de forma similar, al igual que ocurre con la mayoría de los gráficos que se presentan en este apartado.

Figura 5. Pasos a seguir para la generación de un gráfico circular con Minitab

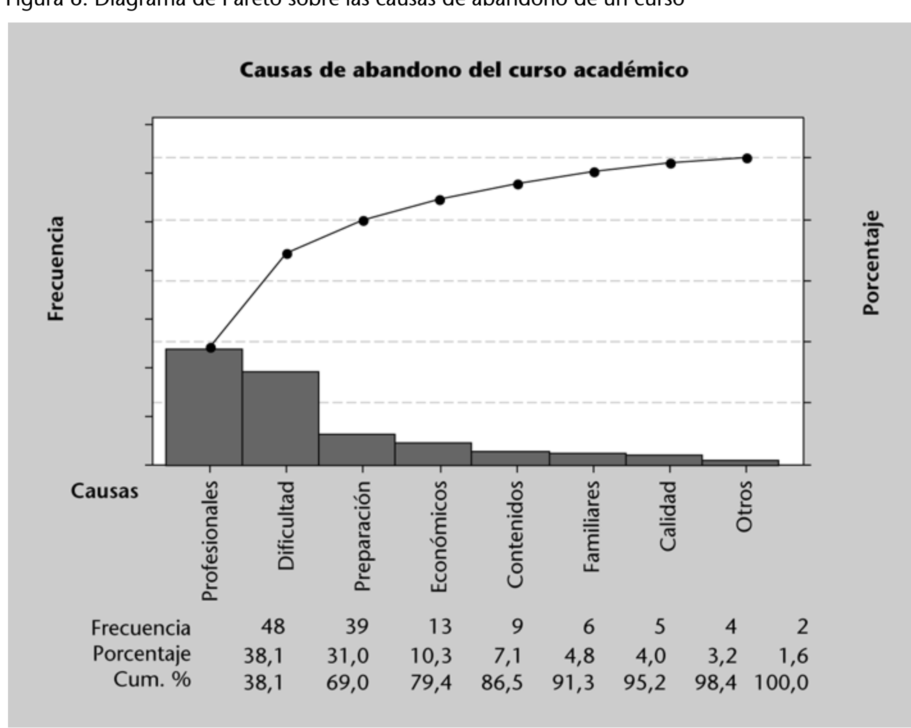


Un gráfico que también suele usarse bastante para describir datos cualitativos es el llamado diagrama de Pareto. Este gráfico está compuesto por: (a) un diagrama de barras en el que las categorías están ordenadas de mayor a menor frecuencia y (b) una línea que representa la frecuencia relativa acumulada (figura 6).

Pasos a seguir

Una vez introducidos los datos en el programa *(1)*, se sigue la ruta ***Graph*** > ***Pie Chart (2)*** y se seleccionan las variables en la ventana correspondiente *(3)*.

Figura 6. Diagrama de Pareto sobre las causas de abandono de un curso



Nota

Las capturas de pantalla de Minitab corresponden a la **versión 15** de este programa. Es posible que otras versiones ofrezcan ligeras diferencias en los menús y ventanas, aunque básicamente el proceso será el mismo. Para obtener más detalles sobre las opciones disponibles, siempre es posible consultar la ayuda en línea del programa o bien alguno de los numerosos manuales de uso que se pueden encontrar en Internet.

Diagrama de Pareto

Para generar un diagrama de Pareto en Minitab hay que usar la ruta *Stat* > *Quality Tools*.

Los diagramas de Pareto son muy útiles para detectar cuándo un porcentaje reducido de categorías (p. ej.: un 20% de las categorías) "acapara" o representa un porcentaje alto de observaciones (p. ej.: un 80% de los datos). Estos fenómenos de excesiva representatividad por parte de unas pocas categorías suelen darse con frecuencia en contextos socioeconómicos (p. ej.: un porcentaje reducido de los ciudadanos de un país acapara un alto porcentaje de la renta), educativos (p. ej.: un porcentaje reducido de causas generan la mayor parte de los abandonos del curso) o de ingeniería de la calidad (p. ej.: un alto porcentaje de fallos son debidos a un número muy reducido de causas). Identificar aquellas pocas categorías que representan una gran parte del porcentaje total puede servir para corroborar ciertos desequilibrios distributivos –como una distribución poco equilibrada de las rentas en un país o de los sueldos en una empresa–, o para proporcionar pistas sobre los principales factores de causa de un problema –como el alto nivel de abandono de un curso o un elevado nivel de fallos en un servicio o producto–.

Gráficos y tablas para datos cuantitativos

En el caso de datos cuantitativos, su representación gráfica o mediante tablas permite apreciar la forma de su distribución estadística, es decir, la forma en que se comporta la variable de interés (cuáles son los valores medios o centrales, cuáles son los valores más habituales, cómo varía, cómo de dispersos son los valores, si muestra algún patrón de comportamiento especial, etc.).

Uno de los gráficos más sencillos de elaborar es el llamado gráfico de puntos (*dotplot*). Se trata de un gráfico en el que cada punto representa una o más observaciones. Los puntos se apilan uno sobre otro cuando se repiten los valores observados (figura 7).

Figura 7. Gráfico de puntos para las calificaciones de un curso



Un gráfico similar, aunque algo más elaborado y con una orientación transpuesta de los ejes, es el llamado diagrama de tallos y hojas (*stem-and-leaf*). En él también se representan los valores observados pero usando los propios valores numéricos en lugar de puntos, lo que proporciona un mayor nivel de detalle. La figura 8 muestra un ejemplo de gráfico de tallos y hojas para los mismos datos empleados en la figura 7. Se observa que el gráfico se ha construido a partir de una muestra de cincuenta calificaciones y que

se ha usado una unidad de hoja (*leaf*) de 0,1. Esto significa que la segunda columna del gráfico representa la parte entera de la calificación, mientras que cada uno de los números situados a su derecha representa la parte decimal de una observación con dicha parte entera. Así, se pueden leer las siguientes calificaciones por orden de menor a mayor: 1,4, 2,9, 3,0, 3,5, 3,9, 4,0, 4,3, etc.

Figura 8. Gráfico de hojas y tallos para las calificaciones de un curso

```
Session

Stem-and-Leaf Display: Calificaciones

Stem-and-leaf of Calificaciones  N  = 50
Leaf Unit = 0.10


 1    1  4
 2    2  9
 5    3  059
 14   4  036788999
(13)  5  0144556677789
 23   6  1223344678
 13   7  00011269
 5    8  2355
 1    9  0
```

Atención

Cabe destacar que en un gráfico de tallos y hojas los datos se apilan de izquierda a derecha en lugar de arriba abajo como ocurre con el gráfico de puntos.

Cuando las observaciones generan un número elevado de valores distintos, resulta recomendable agruparlos en clases o intervalos disjuntos de igual tamaño. De ese modo, cada observación se clasifica en una clase o intervalo según su valor. La tabla 2 muestra un ejemplo de tabla de frecuencias en el que se han agrupado los datos en intervalos. La frecuencia de cada intervalo viene determinada por el número de observaciones cuyos valores están en dicho intervalo. La marca de clase representa el valor medio del intervalo.

Tabla 2. Ejemplo de tabla de frecuencias agrupadas usando intervalos

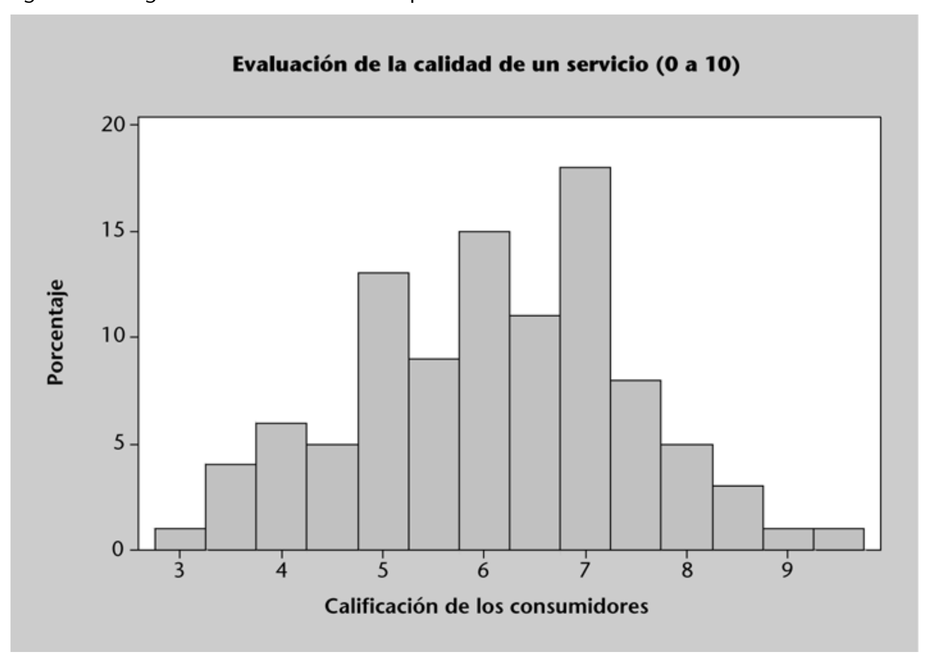
Intervalo	Marca de clase	Frecuencia	Frecuencia relativa
[0, 2)	1	12	8,1%
[2, 4)	3	23	15,5%
[4, 6)	5	67	45,3%
[6, 8)	7	31	20,9%
[8, 10)	9	15	10,1%
Totales		**148**	**100,0%**

Un gráfico que utiliza también intervalos para agrupar los datos a representar es el histograma. El histograma muestra la frecuencia (absoluta o relativa) de cada clase, lo que permite visualizar de forma aproximada la distribución de los datos (figura 9). Sin embargo, hay que tener presente que la forma final del histograma puede variar bastante según el número de intervalos que se definan para agrupar los datos, lo que a veces no permite apreciar correctamente la forma exacta de la distribución estadística que siguen las observaciones.

Nota

Una regla habitual es definir $\sqrt{n}$ clases o intervalos, siendo n el número de observaciones disponibles.

Figura 9. Histograma de una distribución aproximadamente normal



La figura 9 muestra un histograma con forma de campana: es una forma bastante simétrica, que presenta una mayor altura en la parte central y disminuye paulatinamente en las "colas" o extremos. Esta forma es bastante habitual y suele caracterizar el comportamiento de muchas variables (p. ej.: notas numéricas en un examen, peso o altura de individuos, temperaturas diarias, etc.). Sin embargo, también es habitual encontrarse con variables que muestran patrones de comportamientos completamente distintos. Por ejemplo, la figura 10 muestra un histograma en el que se aprecia una distribución más "uniforme" u homogénea de los datos, mientras que la figura 11 muestra un histograma en el que se aprecia una distribución asimétrica o "sesgada" de los mismos.

Figura 10. Histograma de una distribución aproximadamente uniforme

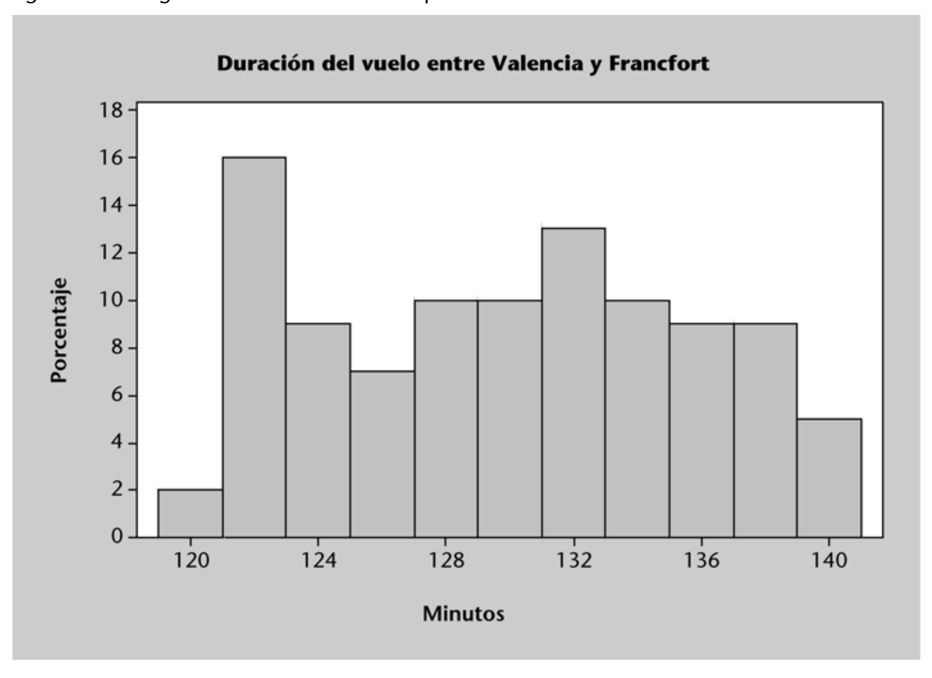
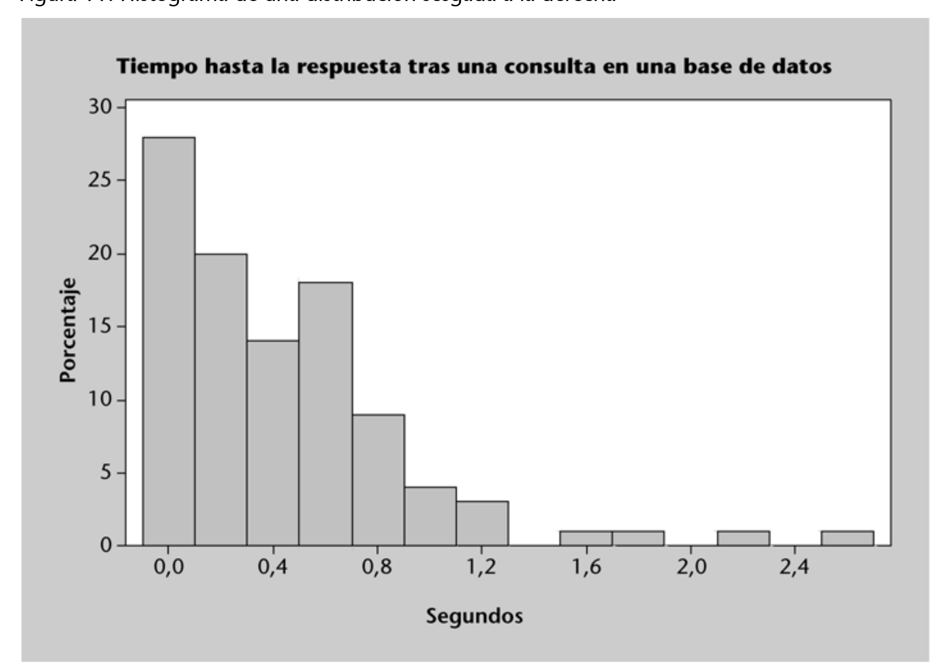


Figura 11. Histograma de una distribución sesgada a la derecha



3. Descripción de datos mediante estadísticos

Dado un conjunto de n datos u observaciones, x_1, x_2, ..., x_n, asociadas a una variable de interés X, suele ser útil sintetizar algunas de sus principales propiedades en unos pocos valores numéricos. Los estadísticos descriptivos son, precisamente, estos valores numéricos capaces de proporcionar información a partir del conjunto de las observaciones. Estos estadísticos resultan muy útiles a la hora de entender el comportamiento de los datos, ya que un simple valor numérico es capaz de describir propiedades tan relevantes como, por ejemplo, el valor promedio del conjunto de datos, el valor máximo, el valor mínimo, el valor que se repite con más frecuencia, un índice de dispersión o variabilidad, etc.

Web

Recordar que la World Wide Web (p. ej., Wikipedia, etc.) es una excelente fuente de consulta para ampliar los conceptos y definiciones estadísticas que se proporcionan en este y otros módulos. Un recurso especialmente interesante, por cuanto ofrece una visión muy completa de conceptos y técnicas estadísticas, es el libro en línea de StatSoft http://www.statsoft.com/textbook/.

Como ya se comentó anteriormente, estos estadísticos hacen referencia a una muestra de observaciones y suelen representarse mediante letras del alfabeto latino ($\bar{x}$, s, etc.), lo que permite distinguirlos claramente de sus parámetros asociados que sintetizan propiedades de toda la población y se representan mediante letras griegas (μ, σ, etc.). Básicamente pueden distinguirse dos grupos de estadísticos descriptivos: (a) los de centralización, que proporcionan información sobre cuáles son los valores "centrales" del conjunto de datos (p. ej.: el valor promedio de los datos) y (b) los de dispersión, que explican cómo se sitúan y varían los datos con respecto a los valores "centrales" (p. ej.: el rango o diferencia entre el valor máximo y el valor mínimo de los datos).

Nota

Recordar que los símbolos μ y σ se pronuncian como "mu" y "sigma", respectivamente. La pronunciación de otros símbolos del alfabeto griego se puede consultar, p. ej., en Wikepedia.

Estadísticos de centralización

A continuación se presentan los estadísticos de centralización más usados habitualmente:

- **Media (*mean*)**: la media (también conocida por valor promedio o valor esperado) de un conjunto de observaciones muestrales se representa con el símbolo $\bar{x}$. Intuitivamente, la media simboliza el "centro de masas" o "punto de equilibrio central" del conjunto de datos considerado. El parámetro asociado, la media poblacional, se representa por μ. Para calcular la media de un conjunto de datos se usa la siguiente expresión:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo: la media de los cinco datos siguientes {6, 3, 8, 6, 4} es $\bar{x} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5,4$

Media muestral

Recordar que la media muestral es un ***estadístico*** que hace referencia al "centro de masas" de los datos de una muestra (subconjunto de la población), mientras que la media poblacional es un **parámetro** que representa el "centro de masas" de toda la población.

- **Mediana (*median*):** la mediana de un conjunto de observaciones muestrales suele representarse con el símbolo $\tilde{x}$. En el caso de una población, el

parámetro mediana se denota con M. Una vez se ordenan todos los datos de menor a mayor, la mediana es aquel valor que deja a su izquierda la mitad de las observaciones (es decir, es aquel valor tal que el número de observaciones más pequeñas que él coincide con el número de observaciones mayores que él). Los pasos para calcular la mediana son: (1) ordenar los datos de menor a mayor, (2) calcular la posición i que ocupa la mediana en el conjunto ordenado de datos, $i = \frac{n+1}{2}$ y (3) seleccionar la observación x_i (la que ocupa la posición determinada en el paso anterior). Cabe observar que si el número de datos n es impar (p. ej.: $n = 5$), la posición i será un valor entero (p. ej.: $i = 3$) que corresponderá con un valor concreto, x_i, del conjunto de datos. Sin embargo, si n es par (p. ej.: $n = 6$), la posición i será un número no entero (p. ej.: $i = 3,5$), en cuyo caso la mediana vendrá dada por el promedio de los dos valores que ocupan las posiciones enteras más cercanas a i (en este caso por el promedio de los valores que ocupan las posiciones 3 y 4).

Ejemplo: dado el conjunto de ocho datos {5, 11, 7, 8, 10, 9, 6, 9}, lo primero es ordenarlos de menor a mayor, con lo que se obtiene la serie {5, 6, 7, 8, 9, 9, 10, 11}; ahora, la posición de la mediana vendrá dada por $i = \frac{8+1}{2} = 4,5$, es decir, la mediana estará entre los valores que ocupan las posiciones 4 y 5, por lo que se calcula el promedio de ambos para dar el valor de la mediana, es decir: $\tilde{x} = \frac{8+9}{2} = 8,5$.

Es importante destacar que la media es muy sensible a la existencia de valores extremos (*outliers*), es decir, la inclusión o no de un valor que esté muy alejado del resto de los datos puede cambiar considerablemente el valor resultante de la media. Por el contrario, la mediana se ve mucho menos afectada por la presencia de dichos valores, lo que significa que la mediana es un "centro" más estable que la media en el sentido de que se ve menos afectado por la presencia de valores extremos en los datos.

- **Moda (*mode*):** la moda de un conjunto de datos es el valor que más veces se repite (el de mayor frecuencia).

  Ejemplo: la moda de la serie de datos {6, 3, 4, 8, 9, 6, 6, 3, 4} es 6, puesto que es el valor que más veces aparece en la serie.

Estadísticos de dispersión

Se presentan ahora los principales estadísticos de dispersión que, como se ha comentado anteriormente, proporcionan información sobre la variabilidad del conjunto de datos:

- **Rango (*range*):** el rango de un conjunto de datos es la diferencia entre el valor máximo y el mínimo de los mismos.

  Ejemplo: dado el conjunto de datos {2, 3, 8, 3, 5, 1, –8}, su rango es 8 – (–8) = 16

- **Varianza muestral (*sample variance*):** la varianza de una muestra se representa por el símbolo s^2. En el caso de una población, el parámetro varianza se representa con el símbolo σ^2. La varianza muestral será mayor cuanto mayor sean las diferencias entre cada una de las observaciones x_i y la media de los datos $\overline{x}$, en concreto:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  Esto significa que la varianza es una medida de la dispersión de los datos con respecto a su media, es decir, cuando menor sea la varianza, tanto más agrupados estarán los datos alrededor de su valor promedio. Por el contrario, cuanto mayor sea la varianza, tanto más dispersos estarán los datos.

  Ejemplo: la varianza muestral de la serie de 5 datos {6, 3, 8, 5, 3} es:

$$s^2 = \frac{(6-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (3-5)^2}{5-1} = 4,5$$

- **Desviación estándar (*standard deviation*):** la desviación estándar (o típica) de una muestra se representa con el símbolo s, mientras que la desviación estándar de una población se representa con σ. La desviación estándar es la raíz cuadrada positiva de la varianza, esto es: $s = \sqrt{s^2}$ (o, dicho de otro modo, la varianza es el cuadrado de la desviación estándar).

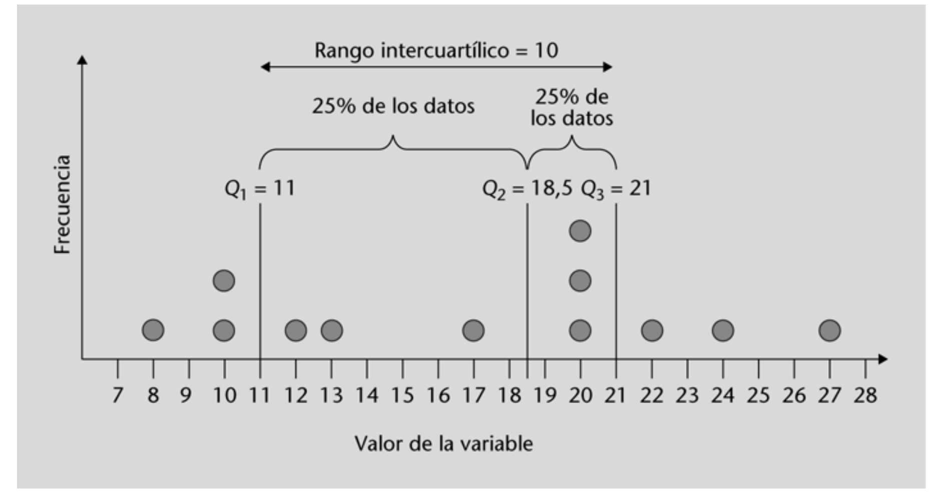
  Ejemplo: para los datos del ejemplo anterior, $s = \sqrt{4,5} = 2,1$

  Al igual que ocurría con la varianza, a mayor desviación estándar más dispersión en los datos y viceversa.

- **Cuartiles (*quartiles*):** en un conjunto de n observaciones ordenadas de menor a mayor valor, se pueden considerar tres valores numéricos concretos llamados cuartiles que dividen el conjunto en cuatro partes, cada una de ellas conteniendo una cuarta parte de las observaciones (figura 12). El primer cuartil, Q_1, es el valor que deja la cuarta parte de los datos ordenados a su izquierda (es decir, un 25% de los datos muestran valores inferiores a él y un 75% de los datos muestran valores superiores a él). Por su parte, el segundo cuartil, Q_2, es aquel valor que deja la mitad de los datos ordenados a su izquierda (es decir, un 50% de los datos muestran valores inferiores a él y un 50% de los datos muestran valores superiores a él). Finalmente, el tercer cuartil, Q_3, es aquel va-

lor que deja tres cuartas partes de los datos ordenados a su izquierda (es decir, un 75% de los datos muestran valores inferiores a él y un 25% de los datos muestran valores superiores a él).

Figura 12. Cuartiles de un conjunto ordenado de datos



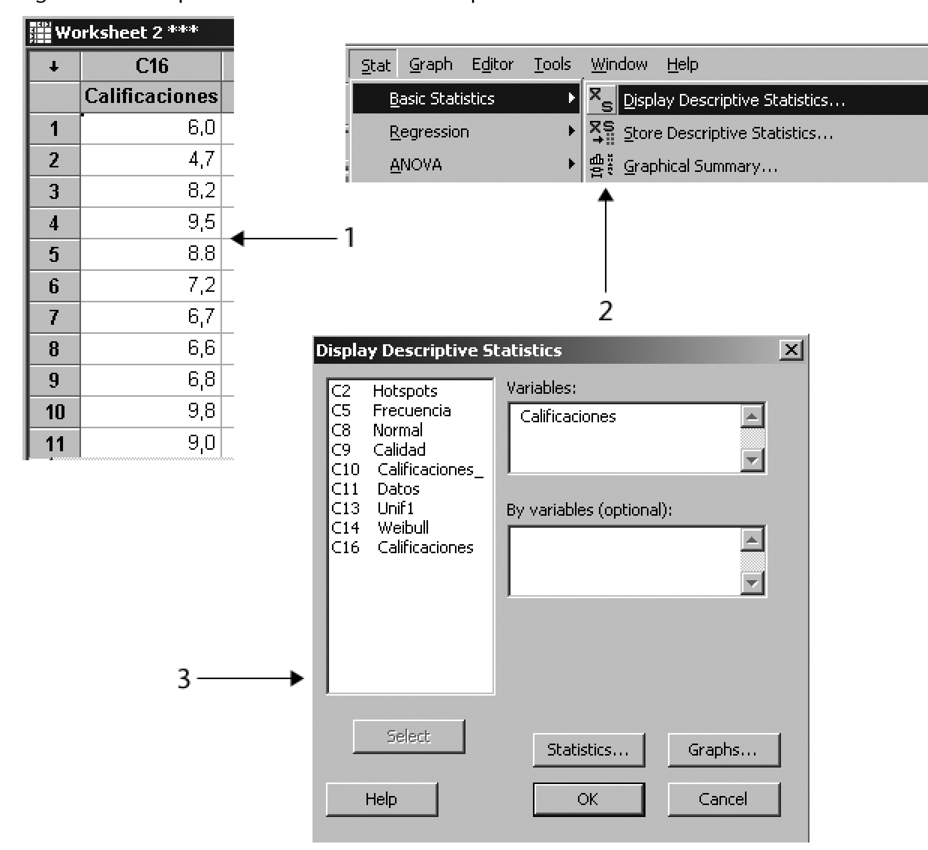
Obsérvese que, en realidad, el cuartil segundo o Q_2 coincide con el concepto de mediana presentado anteriormente. Los cuartiles son muy útiles a la hora de clasificar una observación en una determinada franja del conjunto de datos, por ejemplo, si la observación es inferior a Q_1 significa que ésta se encuentra situada entre el 25% de valores más bajos; si la observación es superior a Q_3 significa que está situada entre el 25% de valores más altos, etc.

- **Rango intercuartílico (*inter-quartilic range*):** este rango suele representarse como *IQR* y es simplemente la diferencia entre el tercer cuartil y el primer cuartil, es decir: $IQR = Q_3 - Q_1$. El rango intercuartílico indica el espacio que ocupan el 50% de las observaciones "centrales" (figura 12), por lo que, de forma similar a lo que ocurría con la varianza, da una medida de la dispersión de los datos (a mayor *IQR* mayor dispersión y viceversa).

Obtención de estadísticos descriptivos mediante programas informáticos

En la práctica, es habitual utilizar algún programa estadístico o de análisis de datos para calcular los estadísticos anteriores e incluso algunos estadísticos adicionales que proporcionen información sobre el conjunto de datos. En la figura 13 se muestran los pasos básicos necesarios para obtener los principales estadísticos descriptivos con Minitab. El *output* del programa, para un ejemplo con cincuenta observaciones, se muestra en la figura 14. Por su parte, la figura 15 muestra una serie de estadísticos descriptivos generados con MS Excel para el mismo conjunto de datos (en este caso los cuartiles se han obtenido usando las fórmulas integradas de Excel).

Figura 13. Pasos para calcular estadísticos descriptivos con Minitab



Pasos a seguir

Una vez introducidos los datos en el programa *(1)*, se sigue la ruta ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics…*** ***(2)*** y se seleccionan las variables en la ventana correspondiente ***(3)***.

Figura 14. Estadísticos descriptivos obtenidos con Minitab

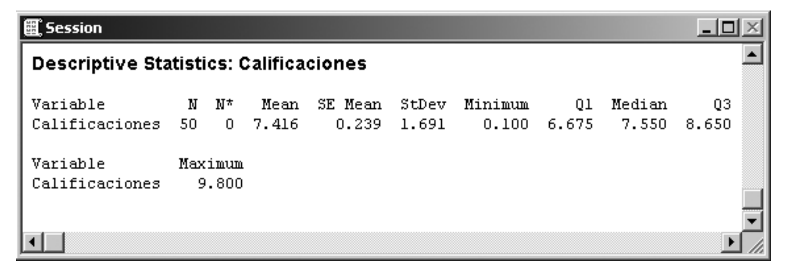


Figura 15. Estadísticos descriptivos calculados con Excel

	A	B	C	D
1	Calificaciones			
2	6,0		*Calificaciones*	
3	4,7			
4	8,2		Media	7,416
5	9,5		Error típico	0,239205488
6	8,8		Mediana	7,55
7	7,2		Moda	7,2
8	6,7		Desviación estándar	1,691438224
9	6,6		Varianza de la muestra	2,860963265
10	6,8		Curtosis	5,922785035
11	9,8		Coeficiente de asimetría	-1,720324942
12	9,0		Rango	9,7
13	7,7		Mínimo	0,1
14	8,6		Máximo	9,8
15	5,8		Suma	370,8
16	6,4		Cuenta	50
17	9,5			
18	7,4		Cuartil primero	6,725
19	7,2		Cuartil segundo	7,55
20	8,8		Cuartil tercero	8,6
21	7,4			

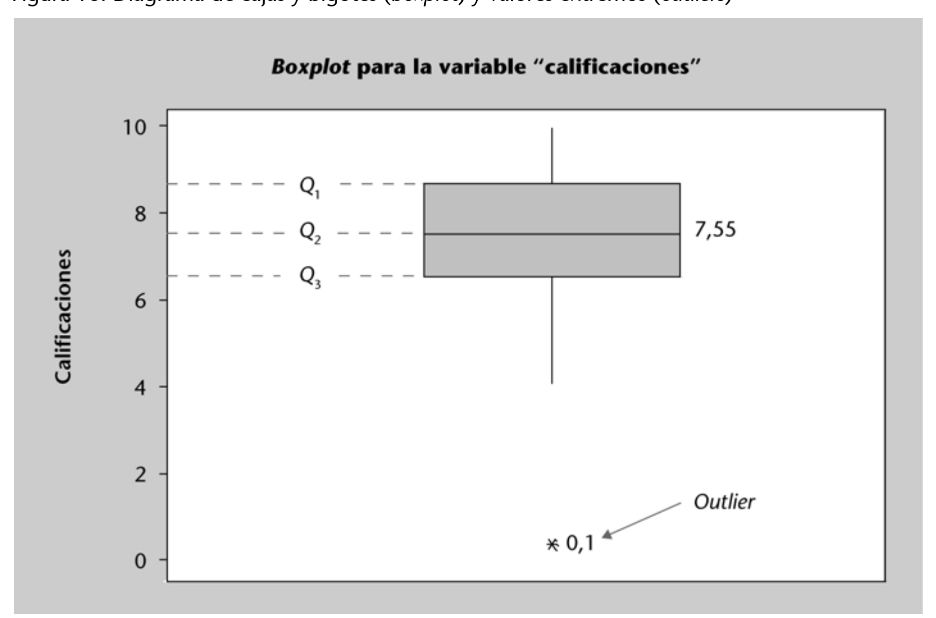
Diferencias en los métodos de cálculos

Cabe destacar que hay ligeras diferencias entre los valores de los cuartiles calculados por Minitab y los correspondientes valores de Excel. Ello se debe a que usan métodos de cálculo distintos. Una discusión interesante sobre los diferentes métodos existentes para calcular los cuartiles se puede encontrar en: http://mathforum.org/library/drmath/view/60969.html.

Diagrama de cajas y bigotes (*boxplot*)

Usando los cuartiles es posible construir un tipo de gráfico, el diagrama de cajas y bigotes (*boxplot*), que resulta muy útil para visualizar la distribución de los datos. Este diagrama está compuesto por una caja central, definida por los cuartiles primero y tercero, que contiene el 50% "central" de las observaciones, y dos segmentos situados en los respectivos extremos de la caja, representando cada uno de ellos el 25% de las observaciones extremas (figura 16).

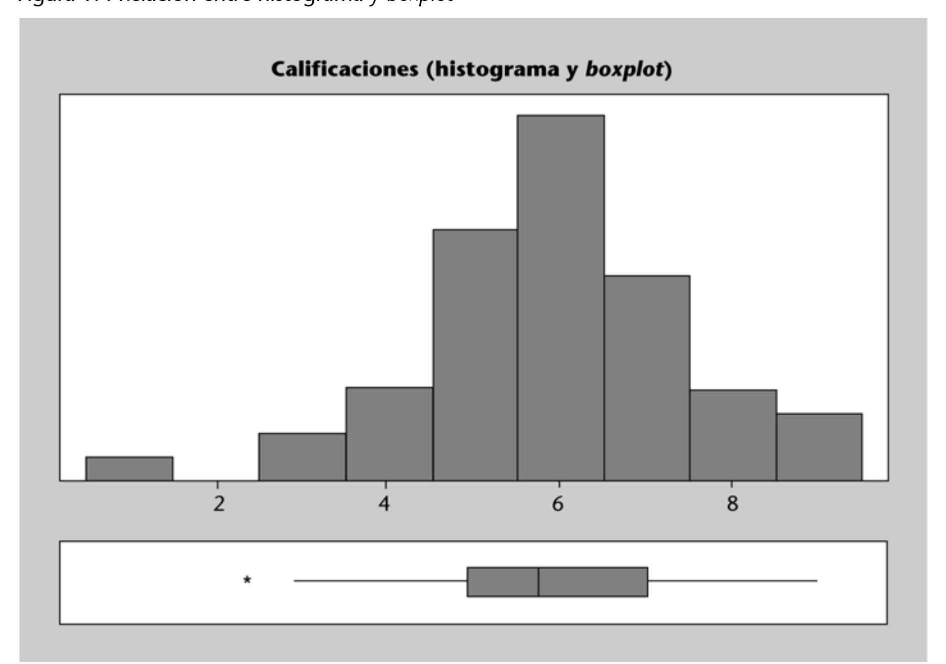
Figura 16. Diagrama de cajas y bigotes (*boxplot*) y valores extremos (*outliers*)



El diagrama de cajas y bigotes sirve también para identificar posibles valores anómalos (*outliers*), que se encuentran excesivamente alejados del resto de los datos, es decir: o bien son extremadamente grandes o bien extremadamente pequeños en comparación con el resto de observaciones. Estos valores anómalos se suelen representar mediante un asterisco, y pueden ser debidos a un error en el registro de los datos o bien a valores que, en realidad, se encuentran extremadamente alejados del resto de observaciones (p. ej.: el precio de un Ferrari cuando se compara con precios de turismos de gama media). Identificar valores anómalos en un conjunto de observaciones es importante, puesto que el análisis de los datos puede dar resultados muy distintos en función de que se consideren o no dichos valores en el estudio (por ejemplo, la media y la varianza de un conjunto de datos pueden cambiar de forma notable según se incluya o no uno de estos valores extremos).

La estrecha relación existente entre el histograma y el *boxplot* se puede observar en la figura 17. En cierto sentido, el *boxplot* se puede interpretar como un histograma visto desde arriba. En este caso, la zona del *boxplot* situada entre los cuartiles primero y tercero correspondería a la zona central del histograma. Además, en ambos casos queda identificado el valor anómalo (*outlier*) así como la forma aproximadamente simétrica del resto de la distribución.

Figura 17. Relación entre histograma y *boxplot*



4. El concepto de probabilidad

Un **experimento aleatorio** es aquel en el que no es posible conocer a priori el suceso resultante que acontecerá pero, sin embargo, sí es posible observar un cierto patrón regular en los resultados que van sucediendo cuando el experimento se repite muchas veces. Por ejemplo, cuando se considera el experimento aleatorio consistente en lanzar una moneda (o un dado) al aire, no es posible predecir cuál será el **suceso resultante** del experimento, es decir, si saldrá cara o cruz (o qué número saldrá en el caso del dado); sin embargo, sí se puede afirmar que tras muchos lanzamientos el porcentaje o proporción de sucesos "cara" obtenidos será muy próximo al 50% o 1/2 (en el caso del dado, el porcentaje o proporción de sucesos "3" obtenidos será muy próximo a 0,1667 o 1/6). Este porcentaje o proporción de aparición de un suceso tras muchas repeticiones del experimento es lo que da lugar a la idea de probabilidad:

Ejemplo

La **probabilidad** de un suceso es siempre un número entre 0 y 1. Así, por ejemplo, una probabilidad de 0,25 representa un porcentaje de aparición del 25% o, equivalentemente, una proporción de 1/4.

> Se define la **probabilidad de un suceso** A, $P(A)$, como el porcentaje o proporción de aparición de dicho suceso en una serie extraordinariamente larga de repeticiones del experimento, todas ellas independientes entre sí.

El requisito de independencia entre las distintas repeticiones del experimento aleatorio significa que el resultado de cada repetición del experimento no está condicionado por los resultados obtenidos en repeticiones anteriores (p. ej.: cuando se lanza varias veces una moneda al aire, el suceso resultante de cada nuevo lanzamiento es independiente de los resultados obtenidos en lanzamientos previos).

Ejemplo 1 de probabilidades

En el experimento "lanzamiento de una moneda al aire", es posible considerar los siguientes sucesos o potenciales resultados: C = {cara}, X = {cruz}, Ω = {cara o cruz} y $\varnothing$ = {ni cara ni cruz}. Los dos últimos sucesos se conocen, respectivamente, como suceso seguro Ω (que incluye todos los resultados posibles) y suceso imposible o conjunto vacío $\varnothing$ (que no incluye ningún resultado derivado de la ejecución del experimento). En este caso, parece claro que $P(C) = 0{,}5$ (es decir, si se repitiera el experimento muchas veces, aproximadamente el 50% de las mismas serían caras), $P(X) = 0{,}5$, $P(\Omega) = 1$ (es decir, en el 100% de los lanzamientos saldrá o bien cara o bien cruz) y $P(\varnothing) = 0$ (es decir, en el 0% de los lanzamientos no se obtendrá resultado alguno).

Ejemplo 2 de probabilidades

En el experimento aleatorio "lanzamiento de un dado", es posible considerar sucesos o potenciales resultados como los siguientes: {1}, {2}, {3}, {4}, {5}, {6},

Ω = {un número entre 1 y 6}, $\varnothing$ = {ningún número entre 1 y 6}. En este caso, $P(\{1\}) = 1/6$ (tras muchas repeticiones, uno de cada seis lanzamientos acabará siendo un 1), $P(\{2\}) = 1/6$, $P(\{3\}) = 1/6$, $P(\{4\}) = 1/6$, $P(\{5\}) = 1/6$, $P(\{6\}) = 1/6$, $P(\Omega) = 1$ y $P(\varnothing) = 0$.

Observar, además, que también es posible considerar sucesos compuestos como, por ejemplo, par = {2, 4, 6}, impar = {1, 3, 5}, mayor2 = {3, 4, 5, 6}, menor3 = {1, 2}, etc. En este caso, *P*(par) = 3/6 = 1/2, *P*(impar) = 1/2, *P*(mayor2) = 4/6 = 2/3, *P*(menor3) = 2/6 = 1/3.

Propiedades básicas de las probabilidades

Hay una serie de propiedades básicas que debe satisfacer cualquier probabilidad. Estas propiedades son muy útiles a la hora de calcular probabilidades de sucesos complejos a partir de probabilidades ya conocidas o fáciles de obtener:

1) La probabilidad de cualquier suceso A siempre es un número situado entre 0 y 1 (ambos inclusive), es decir $0 \leq P(A) \leq 1$.

Ejemplo: en los ejemplos anteriores, todas las probabilidades halladas eran valores entre 0 y 1.

2) La probabilidad del suceso imposible o conjunto vacío $\varnothing$ es siempre 0, es decir, $P(\varnothing) = 0$. En otras palabras, cuando se hace un experimento aleatorio siempre se obtiene algún resultado y, por tanto, la proporción de "no-resultados" es 0.

Ejemplo: en los ejemplos anteriores, $P(\varnothing) = 0$.

3) La suma de las probabilidades de todos los posibles resultados del experimento aleatorio siempre vale 1. En otras palabras, la probabilidad del suceso seguro es siempre 1.

Ejemplo: En el ejemplo de la moneda, $P(\Omega) = 1 = P(C) + P(X)$; en el ejemplo del dado, $P(\Omega) = 1 = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\})$.

4) La probabilidad de que un suceso no ocurra es 1 menos la probabilidad de que sí ocurra, es decir: $P(\textit{no } A) = 1 - P(A)$.

Ejemplo: en el ejemplo de la moneda, $P(C) = 0{,}5 = 1 - P(\textit{no } C) = 1 - P(X)$; en el ejemplo del dado, *P*(par) = 0,5 = 1 – *P*(no par) = 1 – *P*(impar); $P(\varnothing) = 1 - P(\Omega)$.

5) Si dos sucesos *A* y *B* no tienen resultados comunes (son disjuntos), la probabilidad de que ocurra $A \cup B$ es la suma de las probabilidades, es decir, si *A* y *B* son disjuntos, $P(A \cup B) = P(A) + P(B)$.

Ejemplo: en el ejemplo de la moneda, $P(C \cup X) = P(C) + P(X) = 1$; en el ejemplo del dado, $P(\{1, 2\}) = P(\{1\}) + P(\{2\}) = 2/6 = 1/3$; $P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + 0 = 1$.

6) En general, para cualesquiera dos sucesos A y B se cumplirá que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, donde "$A \cap B$" es el conjunto de posibles resultados que satisfacen los sucesos A y B a la vez. Hay que tener en cuenta que cuando A y B son disjuntos (no tienen resultados en común), "$A \cap B$" $= \emptyset$ y, por tanto, $P(A \cup B) = P(A) + P(B) - P(\emptyset) = P(A) + P(B) - 0 = P(A) + P(B)$, que es la expresión vista en la propiedad anterior.

Ejemplo: en el ejemplo del dado, $P(\text{par} \cup \text{mayor2}) = P(\text{par}) + P(\text{mayor2}) - P(\text{par} \cap \text{mayor2}) = 3/6 + 4/6 - 2/6 = 5/6$ (observar que "par $\cap$ mayor2" = {4, 6}).

5. Distribuciones de probabilidad discretas

Al inicio de este módulo se definió el concepto de variable cuantitativa discreta como aquella variable cuantitativa que podía tomar un número finito o contable de valores distintos. Así, un ejemplo de variable discreta sería X = "resultado del lanzamiento de un dado", ya que dicha variable sólo puede tomar seis posibles valores.

Cada uno de los posibles valores de una variable discreta tendrá asociada una probabilidad de ocurrencia (p. ej., en el caso del dado, la probabilidad de obtener un 2 será de 1/6), por lo que parece natural estudiar cómo se distribuyen o comportan dichas probabilidades. En concreto, se puede definir una "función de probabilidad", $f(x)$, que asocie a cada valor x de la variable discreta X su probabilidad de ocurrencia, $P(x)$. Por ejemplo, en el caso de la variable anterior, asociada al experimento aleatorio **"lanzamiento de un dado normal"**, la correspondiente función de probabilidad sería: $f(1) = P(X = 1) = 1/6$, $f(2) = P(X = 2) = 1/6$, $f(3) = P(X = 3) = 1/6$, $f(4) = P(X = 4) = 1/6$, $f(5) = P(X = 5) = 1/6$, $f(6) = P(X = 6) = 1/6$.

Observad

Fijaos que si se usara un **dado "trucado"**, no todas las probabilidades de ocurrencia serían iguales y, por tanto, la función de probabilidad tomaría valores distintos para distintos valores posibles de la variable.

Dada una variable aleatoria discreta X, resulta útil conocer la **distribución de probabilidad** de dicha variable, es decir, cómo se distribuyen o comportan las probabilidades de ocurrencia de sus posibles valores. A tal efecto se definen las siguientes funciones:

La **función de probabilidad** de X es aquella función $f(x)$ que asigna a cada posible valor x de X su probabilidad de ocurrencia, es decir: $f(x) = P(X = x)$ para todo valor posible x de X.

La **función de distribución** de X es aquella función $F(x)$ que asigna a cada posible valor x de X su probabilidad acumulada de ocurrencia, es decir $F(x) = P(X \leq x)$ para todo valor posible x de X.

La tabla 3 muestra la función de probabilidad y la función de distribución correspondientes a la variable X anterior pero usando un dado "trucado" que tiene dos valores 6 y ningún valor 2. Por su parte, la figura 18 muestra ambas funciones superpuestas en el mismo gráfico. Observando detenidamente la tabla 3 y la figura 18 se pueden deducir las siguientes características propias de estas funciones:

- Puesto que representan probabilidades, ambas funciones siempre toman valores en el intervalo [0, 1].

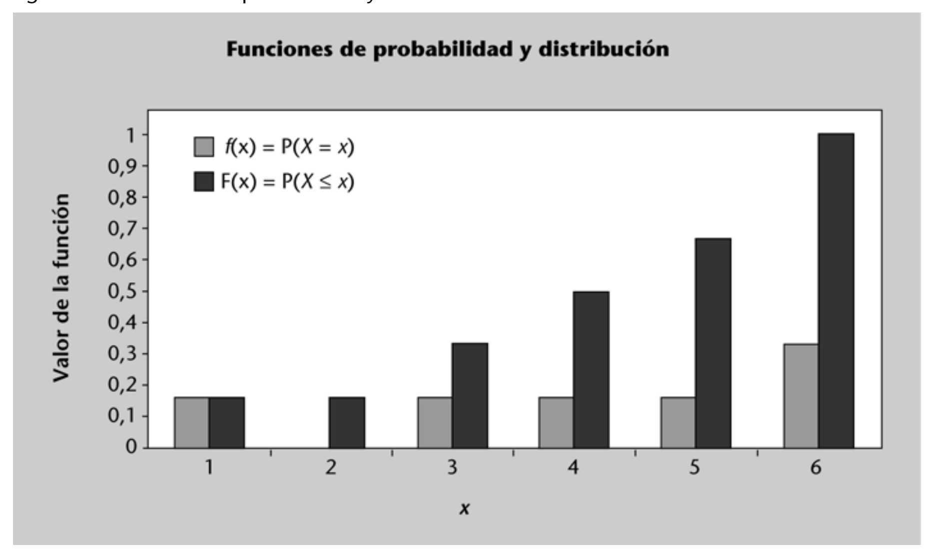
- La suma de todos los valores que toma la función de probabilidad siempre ha de ser 1 (ello se debe a las propiedades de la probabilidad).

La función de distribución siempre es una función creciente que pasa de valor 0 en su extremo izquierdo ($F(0) = P(X \leq 0) = 0$) a valor 1 en su extremo derecho ($F(6) = P(X \leq 6) = 1$).

Tabla 3. Funciones de probabilidad y distribución para una variable discreta

Variable X	Función de probabilidad $f(x) = P(X = x)$	Función de distribución $F(x) = P(X \leq x)$
1	1/6	1/6
2	0	1/6
3	1/6	2/6
4	1/6	3/6
5	1/6	4/6
6	2/6	1
Total	**1**	

Figura 18. Funciones de probabilidad y distribución de una variable discreta



Parámetros descriptivos de una distribución discreta

Mientras que los estadísticos descriptivos y los gráficos o tablas de frecuencias se utilizan para analizar el comportamiento (distribución) de una muestra de observaciones empíricas, las distribuciones de probabilidad son modelos estadísticos que usan parámetros y funciones de distribución para describir el comportamiento teórico (distribución teórica) de toda una población. De forma análoga a lo que ocurría con las muestras –que se caracterizan por estadísticos descriptivos como la media o la varianza muestral–, las distribuciones de probabilidad asociadas a poblaciones también suelen caracterizarse por parámetros tales como la media o la varianza poblacional. Ahora bien, puesto que en general no se dispondrá de observaciones sobre toda la población sino sólo de una función de distribución o de probabilidades, la forma de calcular dichos parámetros es algo distinta:

- **Media o valor esperado de una variable discreta:** la media o valor esperado de una variable discreta X que puede tomar los valores x_1, x_2, ..., se representa con μ o $E[X]$ y se calcula de la siguiente forma:

$$\mu = E[X] = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots = \sum_i x_i \cdot f(x_i)$$

donde $f(x)$ denota a la función de probabilidad de X.

Ejemplo: el caso de un dado equilibrado, el valor esperado o media de X = "resultado del lanzamiento" sería μ = 3; sin embargo, en el caso del dado "trucado" que se muestra en la tabla 3, la media o valor esperado es:

$$\mu = 1 \cdot f(1) + 2 \cdot f(2) + 3 \cdot f(3) + 4 \cdot f(4) + 5 \cdot f(5) + 6 \cdot f(6) =$$
$$= 1 \cdot \frac{1}{6} + 2 \cdot 0 + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{2}{6} = 4{,}167$$

- **Varianza y desviación estándar de una variable discreta:** la varianza de una variable discreta X que puede tomar los valores x_1, x_2, ..., se representa con σ^2 y se calcula de la siguiente forma:

$$\sigma^2 = (x_1 - \mu)^2 \cdot P(X = x_1) + (x_2 - \mu)^2 \cdot P(X = x_2) + \ldots = \sum_i (x_i - \mu)^2 \cdot f(x_i)$$

donde $f(x)$ denota a la función de probabilidad de X. De forma análoga a cómo ocurría con los estadísticos muestrales, la desviación estándar de una variable es la raíz cuadrada positiva de su varianza, es decir:

$$\sigma = \sqrt{\sigma^2}$$

Ejemplo: en el caso del dado "trucado" que se muestra en la tabla 3, la varianza es:

$$\sigma^2 = (1 - 4{,}167)^2 \cdot \frac{1}{6} + (2 - 4{,}167)^2 \cdot 0 + (3 - 4{,}167)^2 \cdot \frac{1}{6} +$$
$$+ (4 - 4{,}167)^2 \cdot \frac{1}{6} + (5 - 4{,}167)^2 \cdot \frac{1}{6} + (6 - 4{,}167)^2 \cdot \frac{2}{6} = 3{,}139$$

Y la correspondiente desviación estándar: $\sigma = \sqrt{3{,}139} = 1{,}772$

La distribución binomial

Una de las distribuciones discretas más usadas en la práctica es la distribución binomial. Esta distribución se usa para contestar a preguntas como las siguientes:

- Si cada vez que un sistema informático es atacado por un virus la probabilidad de que el sistema no falle es de 0,76, ¿cuál es la probabilidad de que no se haya producido ningún fallo en el sistema tras cinco ataques?

- Si cada vez que se consulta una fuente de información la probabilidad de que ésta proporcione una respuesta satisfactoria es de 0,85, ¿cuál es la probabilidad de que se obtenga alguna respuesta satisfactoria tras tres consultas?

- Si tras la administración de un fármaco a un paciente en estado crítico la probabilidad de supervivencia de éste es de 0,99, ¿cuál es la probabilidad de que sobrevivan los catorce pacientes críticos que han recibido el tratamiento?

- Si la probabilidad de obtener una concesión para un proyecto de investigación es de 0,20, ¿cuál es la probabilidad de obtener al menos una concesión tras tres intentos?

- Si cada vez que se trata de encuestar a un transeúnte elegido al azar la probabilidad de que responda es de 0,15, ¿cuál es la probabilidad de que se consigan obtener ochenta respuestas o más a partir de una muestra aleatoria de ciento cincuenta transeúntes?

Distribución de Poisson y la uniforme discreta

Otras distribuciones discretas muy habituales son la distribución de Poisson y la uniforme discreta. Es posible encontrar en Internet abundante documentación sobre éstas y otras distribuciones discretas así como sobre sus ámbitos de aplicación.

La **distribución binomial** es un modelo estadístico que permite calcular probabilidades sobre la variable aleatoria X = "número de éxitos conseguidos en n pruebas independientes". Cada una de estas n pruebas es una repetición de un experimento aleatorio cuyo resultado es binario (éxito o fracaso), siendo p la probabilidad de "éxito" en cada prueba y $q = 1 - p$ la probabilidad de "fracaso".

Resultado "éxito"

No debe confundirse el resultado **"éxito"** de un experimento aleatorio con el hecho de que el resultado sea deseable desde un punto de vista social o subjetivo. Así, por ejemplo, se podría considerar "éxito" del experimento aleatorio el fallo del sistema informático que sufre el ataque de un virus.

Cabe observar que la variable X = "número de éxitos en n pruebas independientes" puede tomar cualquier valor k entre 0 y n (ambos inclusive). Se suele usar la notación $X \sim B(n, p)$ para indicar que X se distribuye o se comporta según una distribución binomial de parámetros n (número de pruebas o repeticiones) y p (probabilidad de "éxito" en cada prueba). En tales condiciones, las probabilidades asociadas a dicha variable vienen dadas por la expresión matemática siguiente:

Para cualquier k entre 0 y n, $P(X = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$, donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$,

siendo $0! = 1! = 1$ y $n! = n \cdot (n-1) \ldots 1$ para todo $n > 1$.

Observad

La expresión "$n!$" se lee como "factorial de n" o "n factorial". Así, por ejemplo, $4! = 4 \cdot 3 \cdot 2 \cdot 1$ y $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. Sin embargo, $1! = 1$ y $0! = 1$.

Se cumple, además, que la media (valor esperado) y la varianza de una distribución binomial son, respectivamente: $\mu = n \cdot p$ y $\sigma^2 = n \cdot p \cdot (1 - p)$.

Ejemplo: la probabilidad de que al introducir datos en un formulario web se cometa un error es de 0,1. Si diez personas rellenan el formulario de forma independiente, ¿cuál es la probabilidad de que no haya más de un formulario erróneo?, ¿cuál es el valor esperado y la desviación estándar de la variable considerada?

Fijémonos en que, en este caso, X = "número de formularios erróneos en diez pruebas" y $X \sim B\,(10, 0{,}1)$. Además, se pide $P(X \leq 1) = P(X = 0 \cup X = 1) = P(X = 0) + P(X = 1)$ (puesto que son sucesos disjuntos). Ahora bien:

$$P(X=0)=\binom{10}{0}0{,}1^{0}\cdot(0{,}9)^{10}=\frac{10!}{0!10!}(1)(0{,}3487)=0{,}3487$$

$$P(X=1)=\binom{10}{1}0{,}1^{1}\cdot(0{,}9)^{9}=\frac{10!}{1!9!}(0{,}1)(0{,}3874)=0{,}3874$$

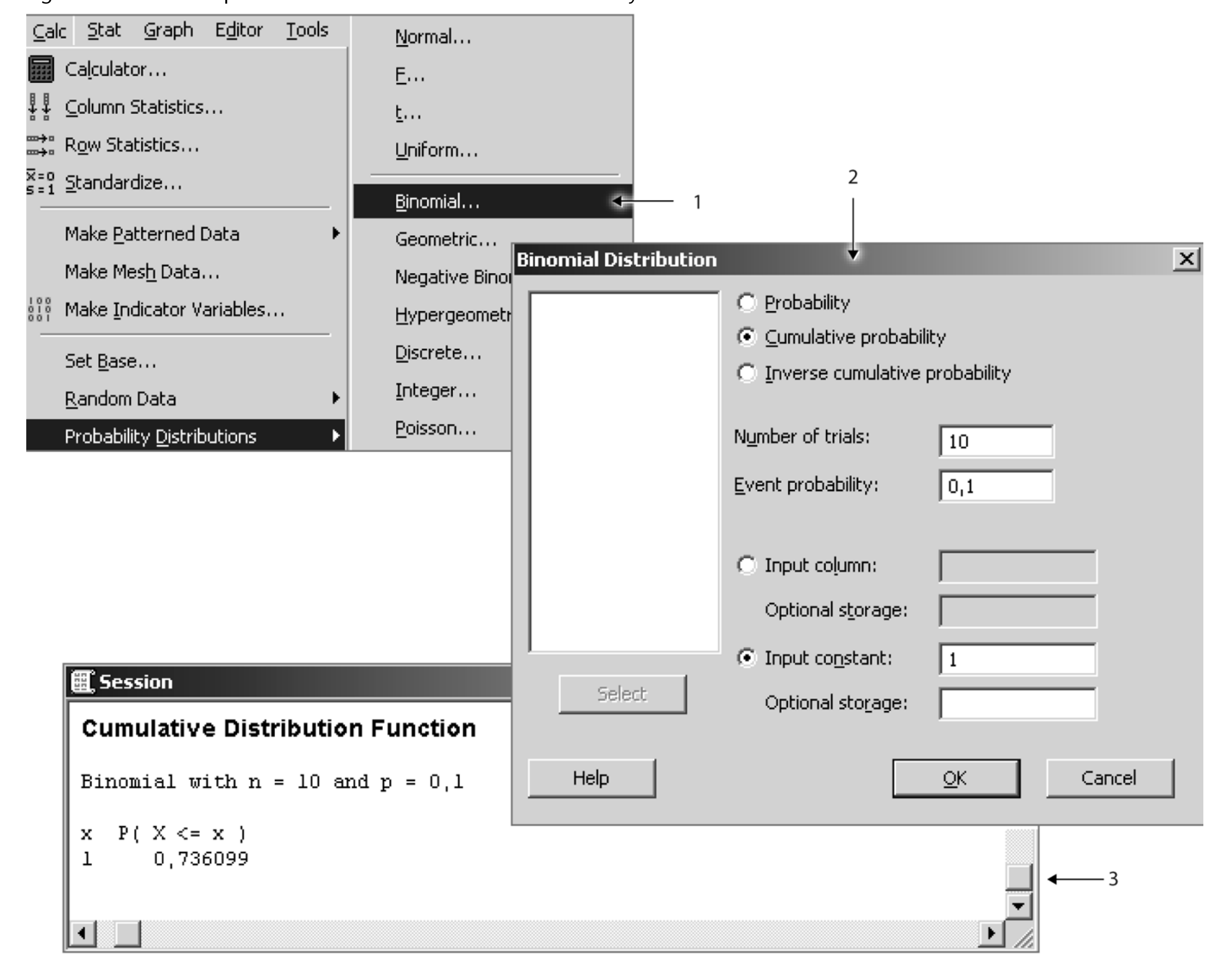
Por tanto, $P(X \leq 1) = 0{,}3874 + 0{,}3487 = 0{,}7361$. Finalmente, $\mu = 10 \cdot 0{,}1 = 1$ y $\sigma = \sqrt{10\cdot 0{,}1\cdot 0{,}9} = 0{,}9487$.

En la práctica, los cálculos probabilísticos anteriores se suelen automatizar con la ayuda de algún programa estadístico o de análisis de datos. La figura 19 muestra cómo se pueden calcular probabilidades de una binomial con ayuda de Minitab. La figura 20, por su parte, muestra cómo obtenerlas usando Excel.

Pasos a seguir

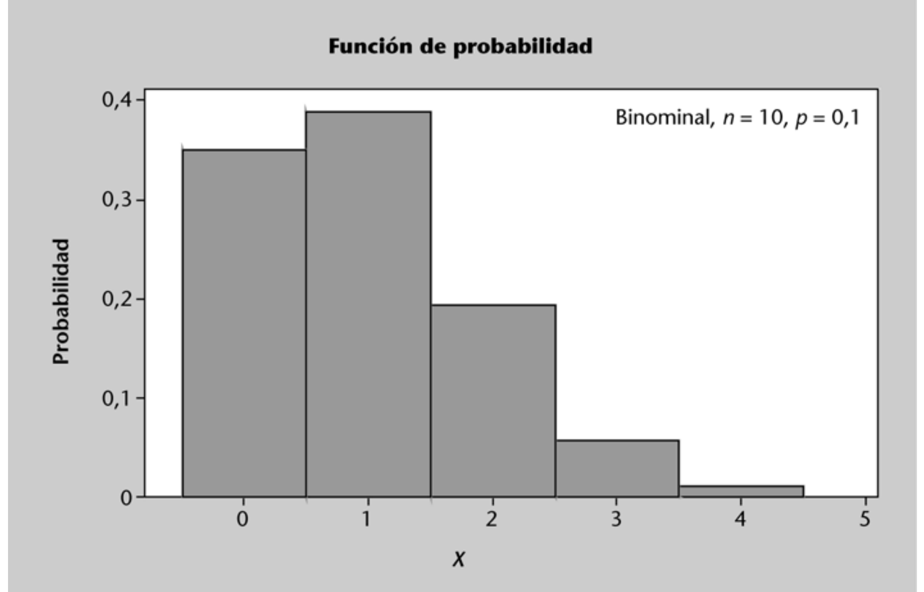
Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***Binomial (1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. Observar que, si en lugar de escoger la opción ***Cumulative probability*** en ***(2)*** se hubiera escogido la opción ***Probability***, el programa hubiera calculado $P(X = 1)$ en lugar de $P(X <= 1)$. Finalmente, para una probabilidad p dada, la opción ***Inverse cumulative probability*** devuelve aquel valor c de la variable X tal que $P(X <= c) = p$.

Figura 19. Cálculo de probabilidades en una binomial con Minitab y Excel



La figura 20 se muestra la función de probabilidad asociada a la binomial del ejemplo anterior. Se observa que, aunque en teoría los posibles valores de la variable X irían desde 0 hasta 10 (número de pruebas), en la práctica los valores mayores de 4 tienen probabilidad de suceso prácticamente nula (por ejemplo, es muy poco frecuente que se obtengan valores superiores a 4). En efecto, $P(X > 4) = 1 - P(X <= 4)$ = {usando Minitab o Excel} = 1 – 0,9984 = 0,0016.

Figura 20. Función de probabilidad de una B (10, 0,1)



Las probabilidades anteriores se pueden obtener también mediante el uso de tablas estadísticas (sin necesidad de usar ningún software). Así, siguiendo el ejemplo anterior, la figura 21 muestra cómo calcular $P(X = 1)$ usando la tabla binomial. En este caso, X es una $B(10, 0,1)$ y se quiere hallar $P(X = k)$ siendo $k = 1$. Para ello, se busca la sección de la tabla correspondiente a $n = 10$, y la intersección entre la fila $k = 1$ y la columna $p = 0,1$.

Cálculo de probabilidades

Resulta fácil encontrar en Internet abundantes documentos que explican con todo detalle el uso de tablas para calcular probabilidades. En la medida de lo posible, sin embargo, conviene automatizar los cálculos mediante el uso de software.

Figura 21. Cálculo de probabilidades binomiales mediante tablas

n	*k*	*p*	0,01	0,05	0,10	0,15	0,20	0,25
	7		0,0000	0,0000	0,0000	0,0000	0,0001	0,0004
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
9	0		0,9135	0,6302	0,3874	0,2316	0,1342	0,0751
	1		0,0830	0,2985	0,3874	0,3679	0,3020	0,2253
	2		0,0034	0,0629	0,0446	0,2597	0,3020	0,3003
	3		0,0001	0,0077	0,0074	0,1069	0,1762	0,2336
	4		0,0000	0,0006	0,0008	0,0283	0,0661	0,1168
	5		0,0000	0,0000	0,0001	0,0050	0,0165	0,0389
	6		0,0000	0,0000	0,0000	0,0006	0,0028	0,0087
	7		0,0000	0,0000	0,0000	0,0000	0,0003	0,0012
	8		0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	9		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
10	0		0,9044	0,5987	0,3487	0,1969	0,1074	0,0563
	1		0,0914	0,3151	0,3874	0,3474	0,2684	0,1877
	2		0,0042	0,0746	0,1937	0,2759	0,3020	0,2816
	3		0,0001	0,0105	0,0574	0,1298	0,2013	0,2503
	4		0,0000	0,0010	0,0112	0,0401	0,0881	0,1460
	5		0,0000	0,0001	0,0015	0,0085	0,0264	0,0584

p

$P(X = 1) = 0,3874$

n

k

6. Distribuciones de probabilidad continuas

Al inicio de este módulo se definió el concepto de variable cuantitativa continua como aquella variable cuantitativa que podía tomar un número infinito (no contable) de valores distintos. Así, un ejemplo de variable continua sería X = "tiempo que se tarda en desarrollar un portal web", ya que esta variable puede tomar un valor real cualquiera entre 0 e infinito.

A diferencia de lo que ocurría con las variables discretas, cuando se trabaja con variables continuas no es posible definir una función de probabilidad que asigne probabilidades a los distintos valores de la variable: si X es una variable continua, X puede tomar un número infinito (no contable) de valores, por lo que la probabilidad teórica de que la variable X tome un valor concreto x es siempre 0, es decir: $P(X = x) = 0$ para cualquier valor x de X. Sí es posible, sin embargo, asignar probabilidades a intervalos de valores. Por ejemplo, si el 51% de los portales web tardan en desarrollarse entre 240 y 258 horas, entonces $P(240 < X < 258) = 0{,}51$. Para describir la distribución de probabilidad de una variable continua se sigue usando la función de distribución (aunque con algún matiz nuevo) y, además, se usa también la llamada "función de densidad" en lugar de la función de probabilidad típica de variables discretas:

Nota

En variables continuas, puesto que $P(X = x) = 0$ para cualquier valor x de X, se cumplirá que:

a) $P(X \leq x) = P(X < x)$

b) $P(X \geq x) = P(X > x)$

La **función de densidad** de una variable continua X es una función $f(x)$ tal que la probabilidad de que X tome un valor en un intervalo (a, b) coincide con el **área** "encerrada" por dicha función entre los extremos de dicho intervalo (figura 22), es decir: $P(a < X < b)$ = área bajo $f(x)$ entre a y b.

La **función de distribución** de X es aquella función $F(x)$ que asigna a cada posible valor x de X su probabilidad acumulada de ocurrencia (figura 23), es decir, $F(x) = P(X \leq x)$ = área bajo $f(x)$ desde $-\infty$ (menos infinito) hasta x.

Nota

La función de densidad $f(x)$ siempre es positiva y "encierra" un área total de 1.

Atención

Observar la equivalencia entre los conceptos de "probabilidad" y "área".

La figura 22 muestra la función de densidad de una variable con distribución simétrica y centrada en el valor 250 (puesto que la función es totalmente simétrica la media y la mediana coinciden en este punto). Se observa también el área encerrada bajo función de densidad entre los valores $a = 240$ y $b = 258$. Esta área corresponde con la probabilidad siguiente: $P(240 < X < 258)$. Por su parte, la figura 23 muestra la función de distribución asociada a la misma variable. Nuevamente se aprecia la simetría con respecto al valor central, así como el hecho de que la función de distribución va creciendo conforme va acumulando probabilidades, pasando del valor 0 en su extremo izquierdo al valor 1 en su extremo derecho. A partir de esta gráfica se pueden estimar visualmente probabilidades acumuladas, por ejemplo: $P(X <= 260)$ será un valor muy cercano a 0,8.

Figura 22. Función de densidad de una variable continua y área encerrada

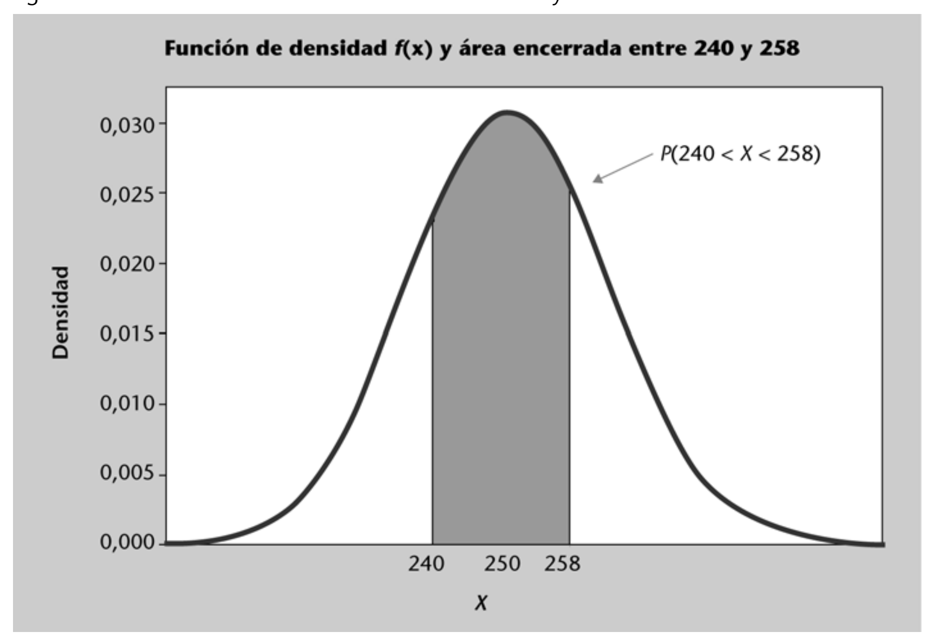
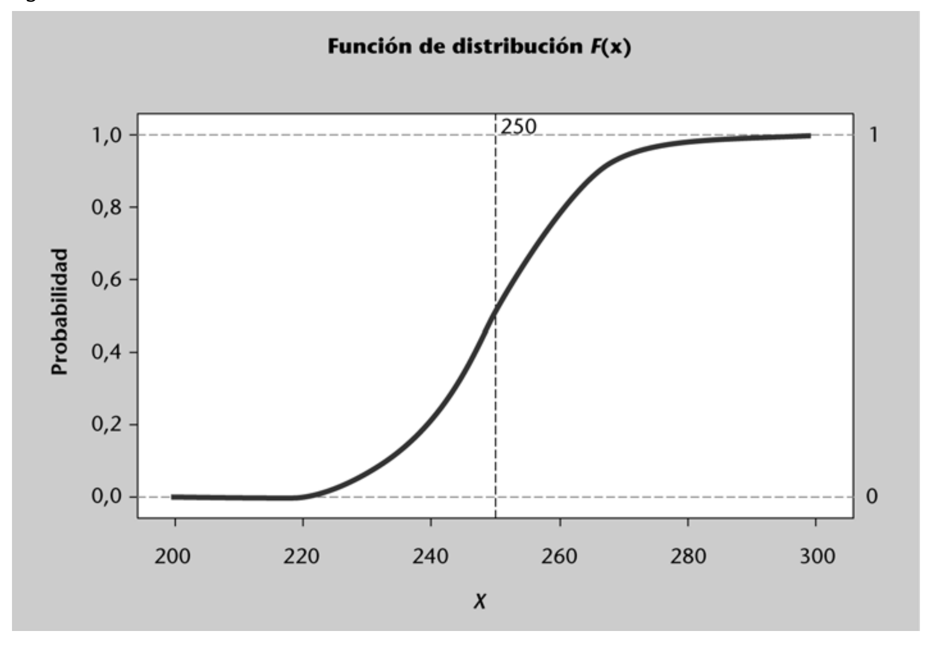


Figura 23. Función de distribución de una variable continua



Función de distribución

La función de distribución es una función acumulativa de probabilidades y, por tanto, es siempre creciente, pasando de 0 (extremo izquierdo) a 1 (extremo derecho).

Parámetros descriptivos de una distribución continua

En el caso de distribuciones continuas, la forma de calcular los parámetros es similar a la empleada para distribuciones discretas, si bien ahora los sumatorios se sustituyen por áreas (integrales definidas en términos matemáticos) entre dos extremos:

Atención

Aunque en la práctica se hará uso de programas estadísticos para hacer los cálculos, es importante conocer qué conceptos se usan para definir cada tipo de parámetro.

- **Media o valor esperado de una variable continua:** la media o valor esperado de una variable continua X se representa por μ o $E[X]$ y se calcula de la siguiente forma:

$$\mu = E[X] = \text{área total bajo “} x \cdot f(x) \text{”} = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

donde $f(x)$ denota a la función de densidad de X.

- **Varianza y desviación estándar de una variable continua:** la varianza de una variable continua X se representa por σ^2 y se calcula de la siguiente forma:

$$\sigma^2 = \text{área total bajo } \text{“}(x - \mu)^2 \cdot f(x)\text{”} = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\,dx$$

donde $f(x)$ denota a la función de densidad de X. Como siempre, la desviación estándar de una variable es la raíz cuadrada positiva de su varianza, es decir:

$$\sigma = \sqrt{\sigma^2}$$

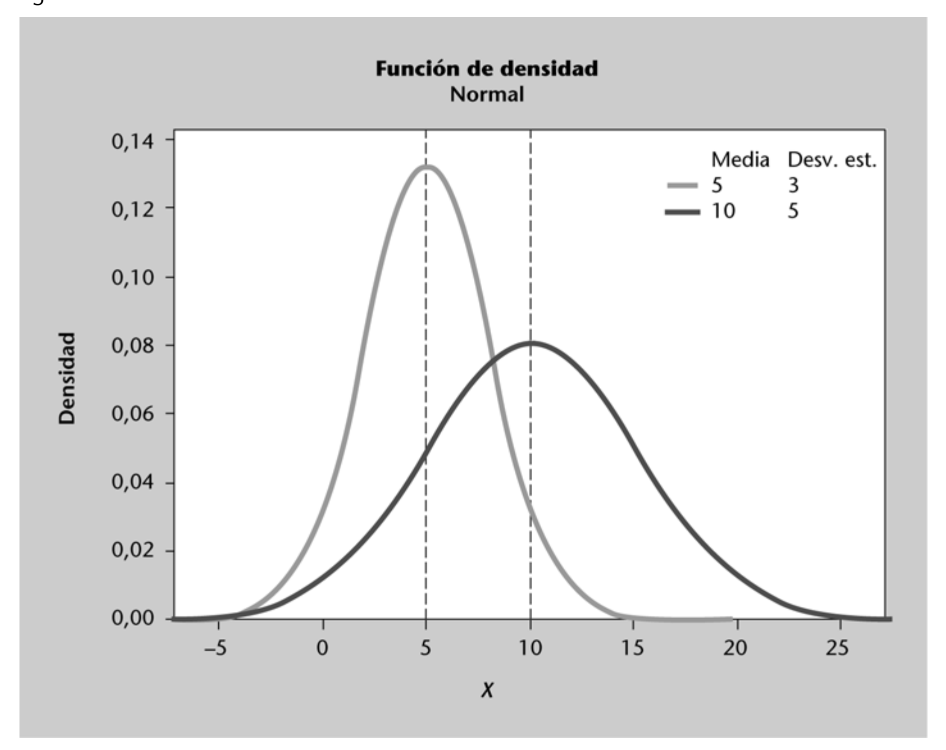
La distribución normal o gaussiana

La distribución normal o gaussiana es la distribución teórica más importante. Muchas variables continuas siguen una distribución normal o aproximadamente normal. Otras variables continuas y discretas también pueden, en determinadas circunstancias, ser aproximadas mediante una distribución normal. La normal, además, es una distribución clave en la estadística inferencial ya que algunas de sus propiedades se utilizan para obtener información sobre toda la población a partir de información sobre una muestra.

La forma concreta de una distribución normal viene caracterizada por dos parámetros: la media, μ, que define dónde se sitúa el centro de la función de densidad, y la desviación estándar, σ, que define la amplitud de la función de densidad. Cuando una variable continua X sigue una distribución normal, se suele representar por $X \sim N(\mu, \sigma)$.

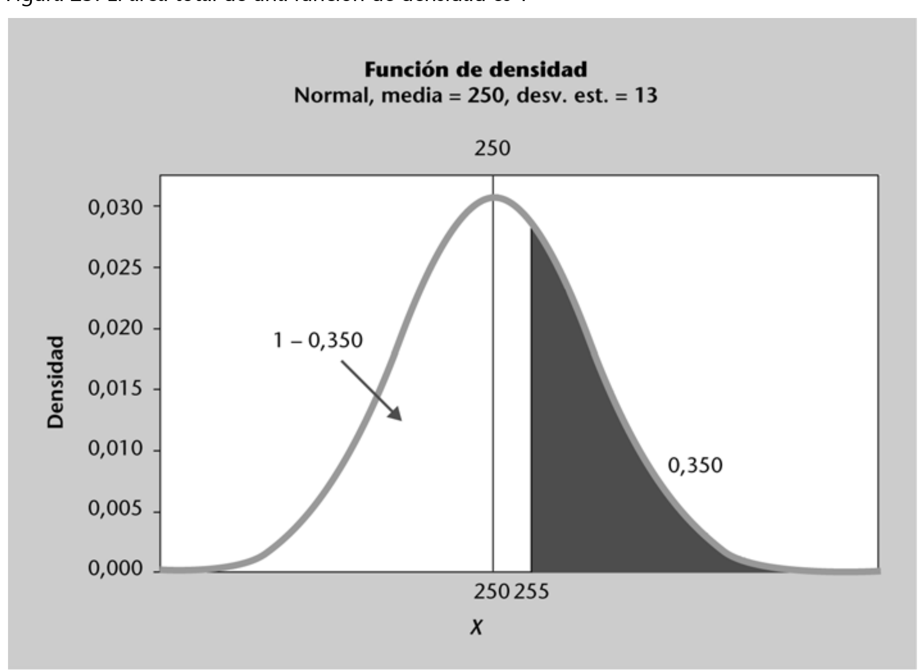
Las figuras 22 y 23 muestran, respectivamente, la función de densidad y la función de distribución de una normal con media $\mu = 250$ y desviación estándar $\sigma = 13$. La figura 24 muestra las funciones de densidad para dos distribuciones de tipo normal con parámetros $\{\mu = 5, \sigma = 3\}$ y $\{\mu = 10, \sigma = 5\}$ respectivamente. Se observa que la función de densidad de la normal tiene forma de “campana de Gauss”, elevada en el centro (el valor medio o esperado) y con dos colas simétricas en los extremos. Es de destacar, además, cómo cada una de las curvas está centrada en su media, así como el hecho de que la curva es más ancha cuanto mayor es la desviación estándar.

Figura 24. Funciones de densidad asociadas a sendas normales



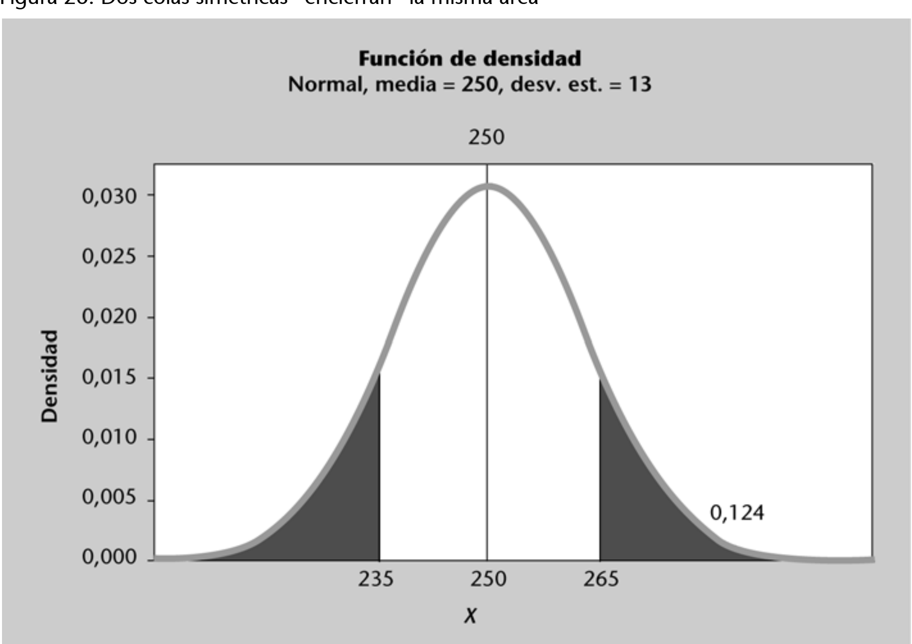
Como en cualquier otra función de densidad, el área total encerrada bajo la curva es de 1. En la práctica eso significa que para cualquier valor x de X, $P(X > x) = 1 - P(X < x)$, es decir, el área a la derecha de un valor es el área total (que vale 1) menos el área a su izquierda y viceversa (figura 25). Además, puesto que la normal es una distribución simétrica con respecto a su media, el área "encerrada" por una cola es igual al área "encerrada" por la cola opuesta (figura 26).

Figura 25. El área total de una función de densidad es 1



Cualquier distribución normal cumple además la llamada **regla 68-95-99,7** según la cual el intervalo ($\mu - \sigma$, $\mu + \sigma$) contiene aproximadamente el 68% de las observaciones, el intervalo ($\mu - 2\sigma$, $\mu + 2\sigma$) contiene aproximadamente el 95% de las observaciones y el intervalo ($\mu - 3\sigma$, $\mu + 3\sigma$) contiene aproximadamente el 99,7% de las observaciones. Así, por ejemplo, si $X \sim N$ (250,13) se puede afirmar que un 68% de las observaciones de *X* estarán en el intervalo (237, 263), un 95% de las observaciones estarán en el intervalo (224, 276) y un 99,7% de las observaciones estarán en el intervalo (211, 289). Observad, por tanto, que será altamente improbable encontrar valores de *X* fuera de este último intervalo.

Figura 26. Dos colas simétricas "encierran" la misma área



De entre las infinitas distribuciones normales que se pueden considerar variando los parámetros μ y σ conviene citar la llamada **normal estándar**, que tiene por parámetros $\mu = 0$ y $\sigma = 1$. En otras palabras, una variable continua *Z* se distribuirá según una normal estándar, $Z \sim N$ (0,1), si su función de densidad es la de una normal centrada en el origen y con desviación estándar unitaria. Esta distribución normal estándar se suele usar bastante en estadística inferencial y también cuando se desean calcular probabilidades de una normal cualquiera mediante el uso de tablas de probabilidades ya calculadas.

En efecto, dada una variable normal cualquiera, $X \sim N$ (μ, σ), es posible aplicarle un **proceso de estandarización** para obtener una normal estándar *Z*. Esto se consigue restando a la variable *X* su media μ (con lo que la función de densidad se desplaza a lo largo del eje *x* hasta que queda centrada en el origen) y dividiendo el resultado por su desviación estándar σ (con lo que la nueva variable tendrá una desviación estándar unitaria), es decir: $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$. Este proceso de estandarización permite, entre otras cosas, calcular probabilidades para una normal cualquiera a partir de las tablas de probabilidades precalculadas que existen para la distribución

normal estándar, lo que evita el tener que resolver integrales cada vez que se desea obtener una nueva probabilidad. Supongamos, por ejemplo, que X sigue una $N(1.500, 100)$ y se desea obtener $P(X < 1.400)$ mediante el uso de tablas. El primer paso consiste en estandarizar los valores:

$$P(X < 1.400) = P\left(\frac{X - \bar{x}}{\sigma} < \frac{1.400 - \bar{x}}{\sigma}\right) = P\left(Z < \frac{1.400 - 1.500}{100}\right) = P(Z < -1)$$

En otras palabras, se desea calcular el área a la izquierda del valor –1 en una normal tipificada o estándar. Normalmente, la tabla de la normal estándar, Z, ofrece áreas (probabilidades) a la izquierda de valores positivos, por lo que resultará necesario hacer una pequeña transformación teniendo en cuenta que: (a) por simetría de la normal estándar, el área (probabilidad) a la izquierda de un valor negativo k es igual al área (probabilidad) a la derecha del correspondiente valor positivo, $|k|$ (p. ej., $P(Z < -1) = P(Z > 1)$), y (b) el área (probabilidad) total encerrada bajo la curva es 1 (p. ej., el área a la izquierda de un valor más el área a su derecha suma 1, por ejemplo: $P(Z < 1) + P(Z > 1) = 1$). Teniendo en cuenta lo anterior, se deduce que $P(Z < -1) = P(Z > 1) = 1 - P(Z < 1)$ = {ver tabla figura 27} = 1 – 0,8413 = 0,1587.

Figura 27. Cálculo de probabilidades en una normal mediante tablas

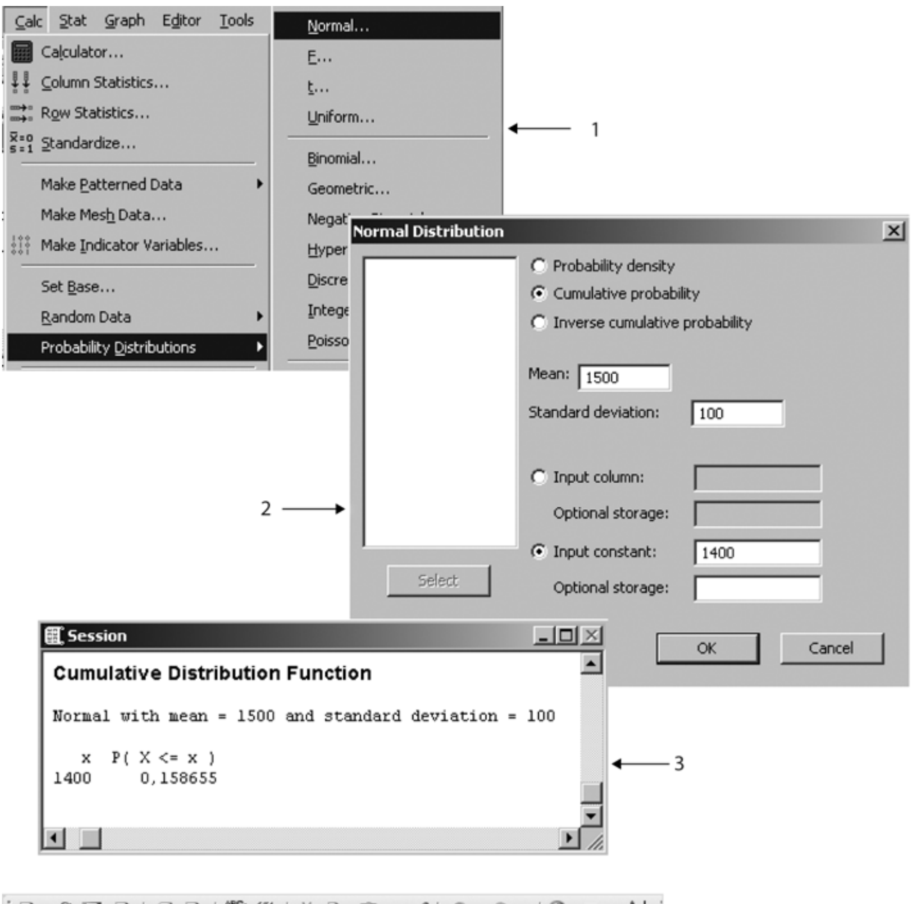
	,00	,01	,02	,03	,04	,05
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265

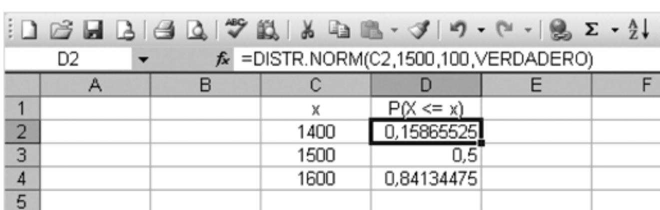
Nota

Notar que para hallar $P(Z < 1,00)$ usando la tabla se ha de buscar el valor intersección entre la fila 1,0 y la columna 0,00 (dado que 1,00 = 1,0 + 0,00). Si se pidiese $P(Z < 1,24)$, entonces habría que buscar la intersección entre la fila 1,2 y la columna 0,04 (dado que 1,24 = 1,2 + 0,04), con lo que se obtendría el valor 0,8925.

Por otra parte, también es posible automatizar el cálculo de probabilidades de una normal cualquiera mediante el uso de programas estadísticos, con lo que se elimina así la necesidad de resolver manualmente las integrales indefinidas o de tener que usar tablas de probabilidades precalculadas. La figura 28 muestra cómo obtener probabilidades de una normal con Minitab. En concreto, para una normal con media $\mu = 1.500$ y desviación estándar $\sigma = 100$, se obtiene que $P(X < 1.400) = 0,158655$. Asimismo, la figura 28 muestra cómo se han obtenido con Minitab y Excel algunas probabilidades para la misma variable. Es preciso observar que $P(X < 1.500) = 0,5$, lo cual es lógico puesto que 1.500 es la media y, a la vez, la mediana de la distribución normal.

Figura 28. Cálculo de probabilidades en una normal con Minitab y Excel





D2 =DISTR.NORM(C2,1500,100,VERDADERO)

	A	B	C	D	E	F
1			x	P(X <= x)		
2			1400	0,15865525		
3			1500	0,5		
4			1600	0,84134475		
5						

Pasos a seguir

Se sigue la ruta ***Calc*** > ***Probability Distributions*** > ***Normal (1)*** y se completan los parámetros en la ventana correspondiente ***(2)***. El resultado se muestra en ***(3)***. Observar que, si en lugar de escoger la opción ***Cumulative probability*** en ***(2)*** se hubiera escogido la opción ***Probability density***, el programa hubiera calculado el valor de la función de densidad en $x = 1.400$ en lugar de $P(X < 1.400)$. Finalmente, para una probabilidad p dada, la opción ***Inverse cumulative probability*** devuelve aquel valor c de la variable X tal que $P(X < c) = p$.

Ejemplos de aplicación de una normal

- Según un estudio realizado por el Ministerio de Educación, el número de horas anuales que dedican los niños españoles a ver la televisión es una variable aleatoria que sigue una distribución normal de media 1.500 horas y desviación estándar de 100 horas. ¿Qué porcentaje de niños dedican entre 1.400 y 1.600 horas anuales?

  En este caso, $X \sim N$ (1.500,100) y se pide $P(1.400 < X < 1.600)$. Por la regla 68-95-99,7, se tiene que la probabilidad anterior será, aproximadamente, del 68% (ya que $\mu - \sigma = 1.400$ y $\mu + \sigma = 1.600$). Para calcular de forma más exacta dicha probabilidad, conviene notar que $P(1.400 < X < 1.600) = P(X < 1.600) - P(X < 1.400)$, es decir: el área entre 1.400 y 1.600 coincide con el área a la izquierda de 1.600 menos el área a la izquierda de 1.400. Las probabilidades anteriores se pueden calcular usando cualquier programa estadístico (p. ej.: Minitab o Excel), y resultan: $P(X < 1.600) = 0,8413$ y $P(X < 1.400) = 0,1587$, por lo que la probabilidad buscada es de 0,6827, es decir, un 68,27% de los niños dedican entre 1.400 y 1.600 horas anuales a ver la televisión.

- En base a los datos del Instituto Nacional de Estadística (INE), el sueldo medio anual de un trabajador es de 26.362 euros. Suponiendo que dichos sueldos sigan una distribución normal con una desviación estándar de 6.500 euros, ¿cuál será el porcentaje de trabajadores que superen los 40.000 euros?

  En este caso, $X \sim N$ (26.362,6.500) y se pide $P(X > 40.000)$. Observar que, puesto que el área total bajo la curva normal es 1, $P(X > 40.000) = 1 - P(X < 40.000)$ = {Minitab o Excel} = 1 – 0,9821 = 0,0179, es decir, sólo un 1,8% de los trabajadores superarían la cifra de los 40.000 euros anuales.

- El tiempo que se emplea en rellenar un cuestionario en línea sigue una distribución aproximadamente normal con una media de 3,7 minutos y una desviación estándar de 1,4 minutos. ¿Cuál es la probabilidad de que se tarde menos de 2 minutos en responder a dicho cuestionario? ¿Y de que se tarde más de 6 minutos? Hallad el valor c tal que $P(X < c) = 0,75$ (percentil 75 de la variable).

  En este caso, $X \sim N$ (3,7, 1,4). En primer lugar, $P(X < 2)$ = {Minitab o Excel} = 0,1131, es decir: un 11,31% de los individuos que respondan el cuestionario emplearan menos de 2 minutos en hacerlo. Por otra parte, $P(X > 6) = 1 - P(X < 6)$ = {Minitab o Excel} = 0,0505, es decir, un 5% de los individuos tardarán más de 6 minutos en responder el cuestionario. Finalmente, para hallar el valor c tal que $P(X < c) = 0,75$ se debe usar la opción *Inverse cumulative probability* de Minitab (o su equivalente en Excel), con lo que se obtiene un valor aproximado de 4,64 minutos, es decir el 75% de los individuos tardan menos de 4,64 minutos en completar el cuestionario (o, dicho de otro modo, el 25% tardan más de 4,64 minutos en hacerlo).

Las distribuciones *t*-Student y *F*-Snedecor

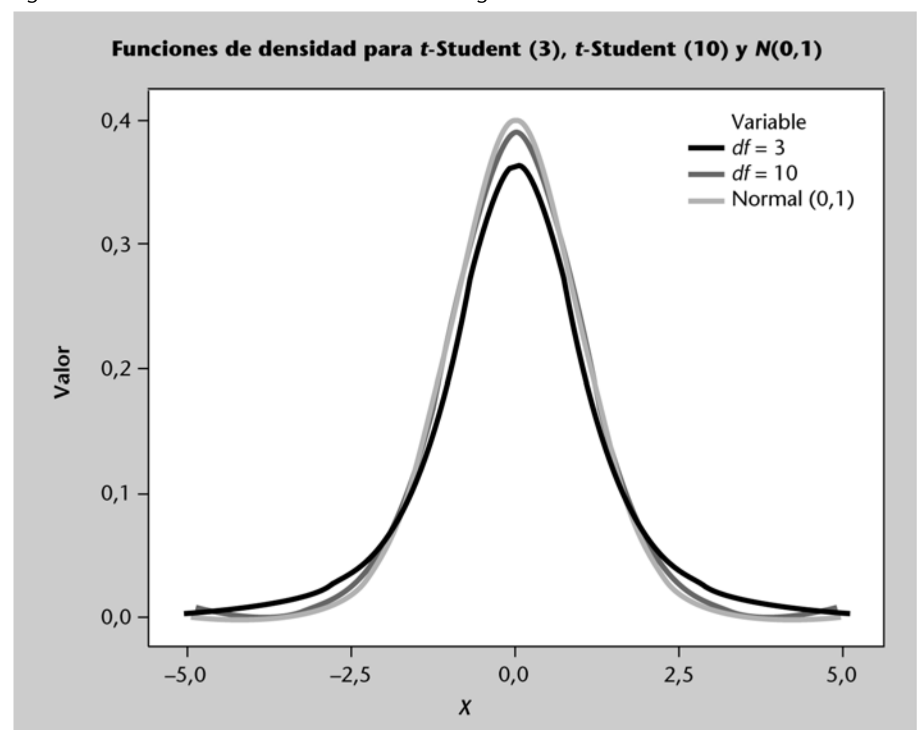
Además de la normal, hay muchas otras distribuciones de probabilidad continuas que se suelen usar en estadística inferencial. Una de ellas es la llamada distribución *t*-Student, y otra es la llamada *F*-Snedecor. Ambas se presentan a continuación:

La distribución ***t*-Student** es una distribución simétrica y centrada en el origen (es decir, su media y su mediana son 0). Esta distribución se caracteriza por un parámetro llamado **grados de libertad** o ***df*** (*degrees of freedom*), siendo $df > 2$. En la práctica, $df = n - 1$, donde n es el tamaño de la muestra que se esté analizando. La figura 29 muestra diversas funciones de densidad de las *t*-Student, cada una de ellas asociadas a un valor concreto del parámetro *df*. Se observa cómo la *t*-Student se asemeja cada vez más a una normal estándar conforme se va incrementando el parámetro grados de libertad.

Grados de libertad

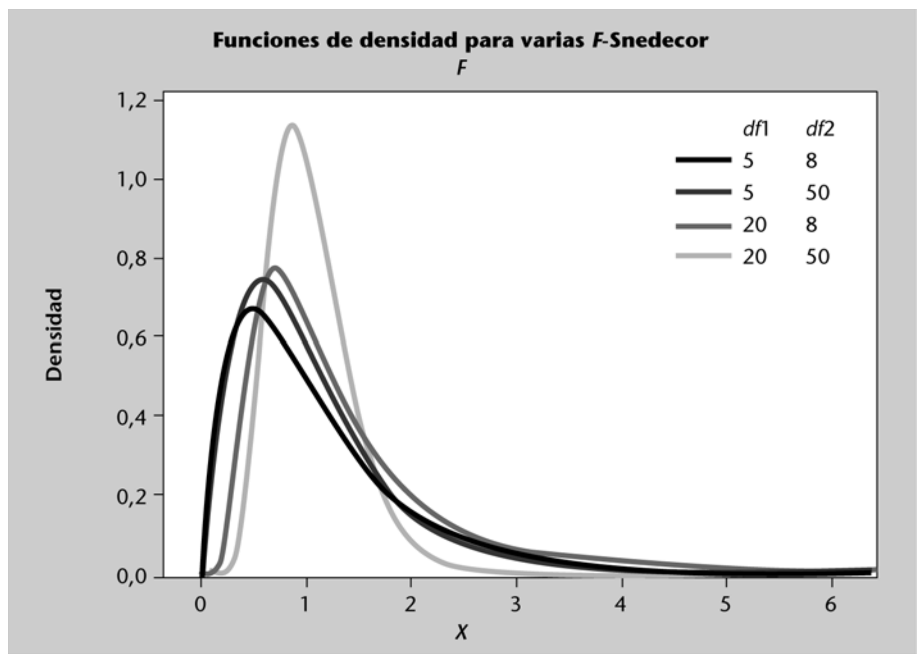
En estadística, el concepto de **grados de libertad** asociados a un conjunto de datos se puede interpretar como el número mínimo de valores que se necesitaría conocer para determinar dichos datos. Así, por ejemplo, en el caso de un muestra aleatoria de tamaño N, habría N grados de libertad (no se puede determinar el valor de ninguno de los datos incluso aunque se conociese el valor de los $N - 1$ restantes). Sin embargo, un conjunto de N datos de los cuales se conozcan $N - 1$, la media muestral tendría $N - 1$ grados de libertad (fijados los valores de los $N - 1$ datos y de la media, quedaría ya fijado el valor desconocido restante). Así, si tenemos un conjunto de 3 observaciones de la variable X, $x_1 = 2$, $x_2 = -2$ y $x_3 = a$ (desconocido), y sabemos que la media de los tres valores es 0, necesariamente $a = 0$.

Figura 29. Funciones de densidad de *t*-Student según *df*



Por su parte, la distribución ***F*-Snedecor** es otra distribución continua. La *F*-Snedecor siempre toma valores no negativos (es decir, una variable que siga dicha distribución sólo puede tomar valores iguales o mayores a 0, nunca valores negativos). Además, esta distribución no es simétrica, sino que está sesgada a la derecha (figura 30). Así como la normal venía caracterizada por dos parámetros, μ (media) y σ (desviación estándar), la *F*-Snedecor también se caracteriza por dos parámetros: los **grados de libertad del numerador**, *df*1 y los **grados de libertad del denominador**, *df*2. Al igual que ocurría con la *t*-Student, para cada valor de estos parámetros se obtiene una función de densidad distinta y, por tanto, una distribución *F*-Snedecor distinta.

Figura 30. Funciones de densidad de *t*-Student según *df*1 y *df*2

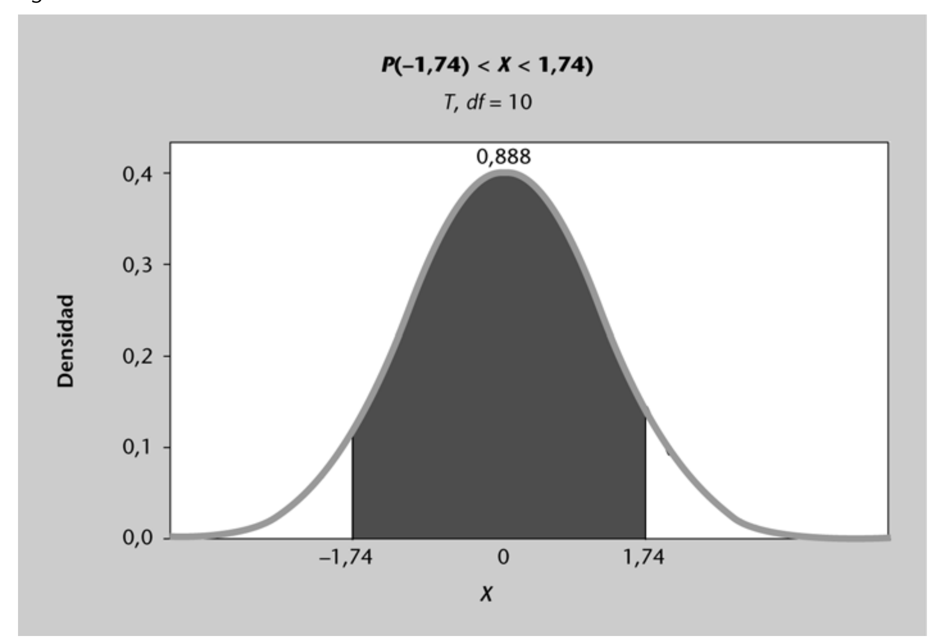


Para calcular probabilidades asociadas a una *t*-Student o a una *F*-Snedecor, pueden usarse programas estadísticos o de análisis de datos (Minitab, Excel, etc.) de forma análoga a como se hacía en el caso de la normal. Así, por ejemplo, si X es una variable aleatoria que sigue una distribución *t*-Student con diez grados de libertad, $P(-1{,}74 < X < 1{,}74) = P(X < 1{,}74) - P(X < -1{,}74)$ = {Minitab o Excel} = 0,9438 – 0,0562 = 0,8876 (figura 31).

Nota

Notar que $P(-1{,}74 < X < 1{,}74)$ viene representada por el área marcada en la figura 31 (esto es, el área comprendida entre los valores –1,74 y 1,74). Para calcular dicha área, se calcula $P(X < 1{,}74)$ (p. ej., el área a la izquierda del 1,74) y al valor obtenido se le resta $P(X < -1{,}74)$ (p. ej., el área a la izquierda del -1,74). Para calcular $P(X < 1{,}74)$ con Minitab se usa el menú ***Calc*** > ***Probability Distributions*** > ***t…***, especificando los grados de libertad (10 en este ejemplo) y el valor de la constante (1,74 en este caso). Análogamente se obtendría el valor de $P(X < -1{,}74)$.

Figura 31. Probabilidades en una *t*-Student

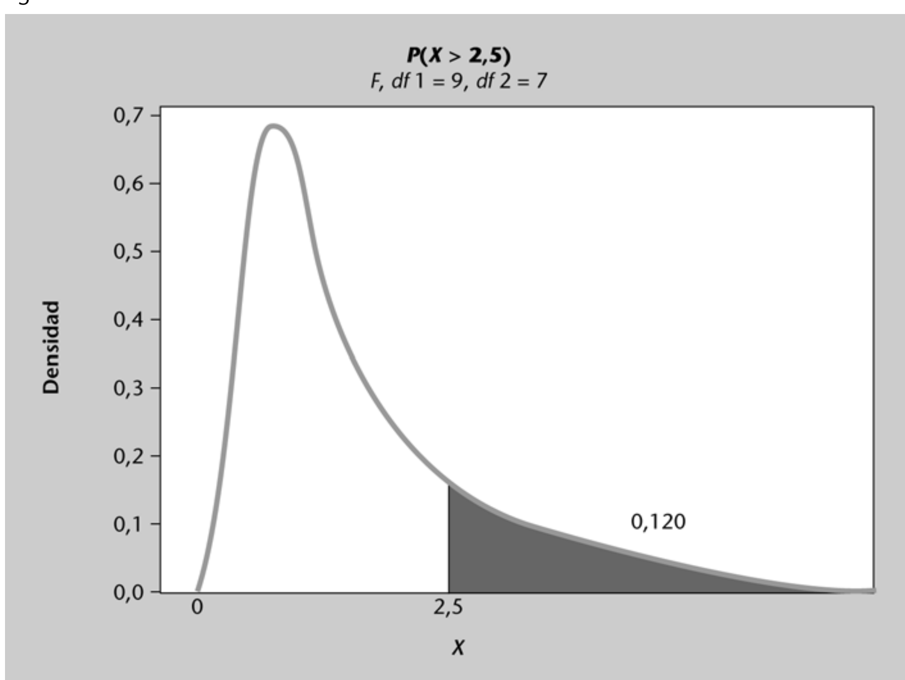


Finalmente, si X es una variable aleatoria que sigue una distribución *F*-Snedecor con nueve grados de libertad en el numerador y siete grados de libertad en el denominador, entonces $P(X > 2{,}5) = 1 - P(X < 2{,}5)$ = {Minitab o Excel} = 1 – 0,8797 = 0,1203 (figura 32).

Nota

De forma análoga a como ocurría en el caso de las distribuciones binomial y normal, también existen tablas que permiten calcular, sin necesidad de utilizar software como Minitab o Excel, las probabilidades asociadas a una distribución *t*-Student o *F*-Snedecor (ver, p. ej., http://www.statsoft.com/textbook/distribution-tables).

Figura 32. Probabilidades en una *F*-Snedecor



Resumen

En este módulo se han presentado las técnicas básicas de la estadística descriptiva univariante: representación gráfica de datos discretos y continuos, organización de los datos mediante tablas de frecuencias y uso de estadísticos descriptivos para resumir datos. Conviene recordar que el tipo de gráfico, tabla o estadístico a usar dependerá siempre del tipo de variable considerada (categórica, cuantitativa discreta o cuantitativa continua), así como del tipo de información que se desee obtener.

Además, se ha explicado también el concepto de probabilidad de un suceso, que desempeña una función relevante en el análisis y predicción del comportamiento de las variables aleatorias asociadas a fenómenos cotidianos.

Finalmente, se han presentado algunos de los principales modelos matemáticos que se usan para describir, de forma teórica, el comportamiento de variables aleatorias: la distribución binomial, la normal, la *t*-Student y la *F*-Snedecor son algunos ejemplos de dichos modelos. El cálculo de probabilidades asociadas a variables que se comportan según alguno de estos modelos permite entender mejor su comportamiento y realizar estimaciones sobre la población de individuos de la que provienen los datos.

Ejercicios de autoevaluación

1) La tabla siguiente resume las respuestas ofrecidas por doscientos usuarios de un portal web a la pregunta "el nivel de usabilidad del portal es adecuado":

Respuesta	Frecuencia
Totalmente de acuerdo	50
De acuerdo	75
Ligeramente de acuerdo	25
Ligeramente en desacuerdo	15
En desacuerdo	15
Totalmente en desacuerdo	20

Se pide que hagáis lo siguiente:
a) Construir un diagrama de barras que permita visualizar las respuestas obtenidas.
b) Calcular la frecuencia relativa de aparición de cada respuesta y construir un diagrama circular para ilustrar dichos valores.

2) La tabla siguiente contiene cuarenta observaciones para el tiempo transcurrido (en horas) entre el envío de un mensaje a un foro en línea y su correspondiente respuesta.

4,0	3,5	3,1	6,0	5,6	3,1	2,9	3,8
4,3	3,8	4,5	3,5	4,5	6,1	2,8	5,0
5,4	3,8	6,8	4,9	3,6	3,6	3,8	3,7
4,1	2,0	3,7	5,7	7,8	4,6	4,8	2,8
5,0	5,2	4,0	5,4	4,6	3,8	4,0	2,9

A partir de estos datos, debéis hacer lo siguiente:
a) Construir un diagrama de tallos y hojas. Usad 1,0 como unidad de incremento.
b) Construir un histograma.
c) ¿Se observa en los datos algún patrón claro? ¿Cuál es la moda de la distribución de los datos?

3) La tabla siguiente muestra veinte observaciones de la variable aleatoria "número de correos electrónicos recibidos en un día".

3,9	3,4	5,1	2,7	4,4
7,0	5,6	2,6	4,8	5,6
7,0	4,8	5,0	6,8	4,8
3,7	5,8	3,6	4,0	5,6

Se pide que hagáis lo siguiente:
a) Hallar los estadísticos descriptivos de esta muestra. ¿Cuánto vale el rango intercuartílico? ¿Entre qué dos valores están comprendidos el 50% de los datos centrales de la muestra?
b) Construir un diagrama de cajas y bigotes (*boxplot*). ¿Hay algún valor anómalo (*outlier*) entre las observaciones?

4) Cuando se efectúa un control antidopaje a un atleta que no ha tomado sustancia alguna, la probabilidad de que el test dé un falso positivo es de 0,006. Si durante una competición se efectúa el test a un total de 1.000 atletas que están libres de sustancias, ¿cuál será el número esperado (promedio) de falsos positivos?, ¿cuál es la probabilidad de que el número de falsos positivos sea superior a quince?, ¿qué cabría pensar si aparecen más de quince positivos?

5) De acuerdo con el Instituto Nacional de Estadística, el 9,96% de los adultos residentes en España son extranjeros. Con el fin de realizar una encuesta, se pretende contactar con una muestra aleatoria de mil doscientos adultos residentes en España. ¿Cuál será el número espe-

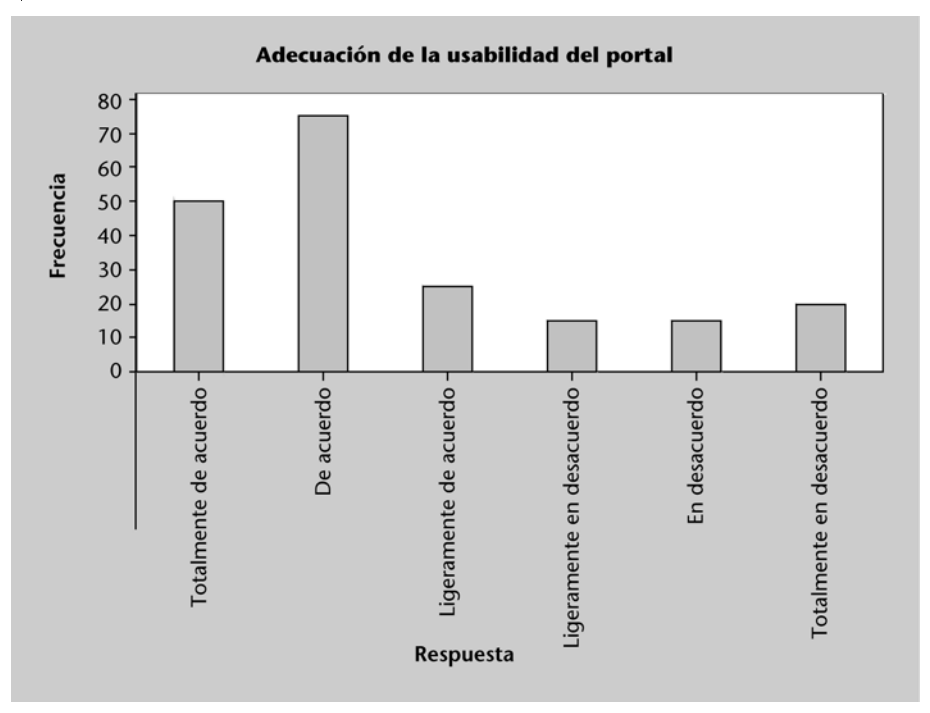
rado (promedio) de extranjeros que contendrá dicha muestra?, ¿cuál es la probabilidad de que la muestra contenga menos de cien extranjeros?

6) El tiempo de duración de un embarazo es una variable aleatoria que se distribuye de forma aproximadamente normal con una media de doscientos sesenta y seis días y una desviación estándar de dieciséis días. ¿Qué porcentaje de embarazos duran menos de doscientos cuarenta días (unos ocho meses)?, ¿qué porcentaje de embarazos duran entre doscientos cuarenta y doscientos setenta días (entre unos ocho y nueve meses)?, ¿a partir de cuántos días se sitúan el 20% de los embarazos más largos?

Solucionario

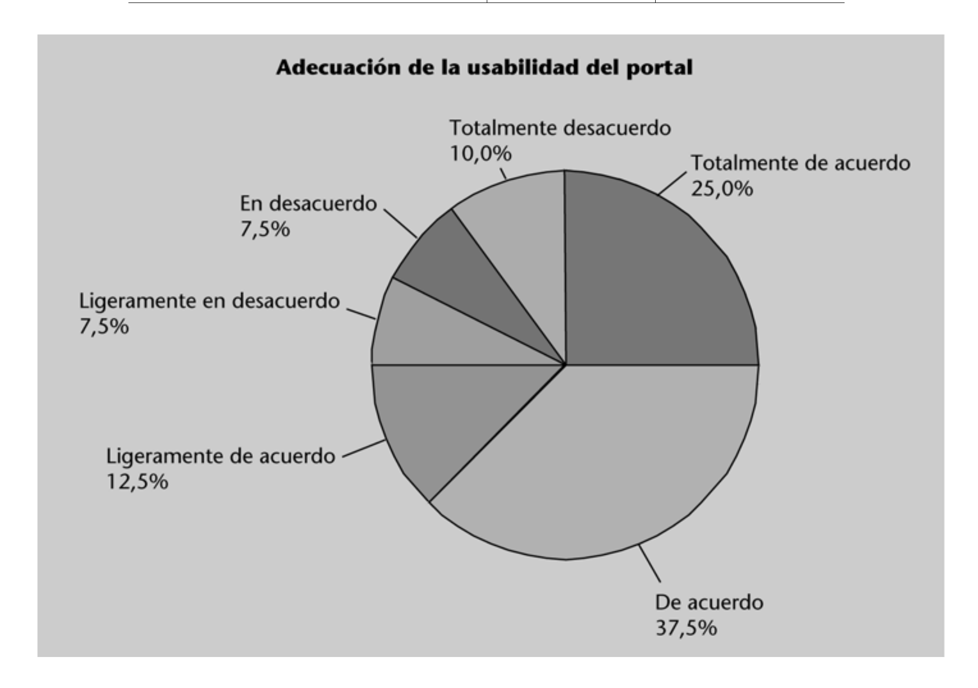
1)

a)



b)

Respuesta	Frecuencia	Frec. relativa
Totalmente de acuerdo	50	25,0%
De acuerdo	75	37,5%
Ligeramente de acuerdo	25	12,5%
Ligeramente en desacuerdo	15	7,5%
En desacuerdo	15	7,5%
Totalmente en desacuerdo	20	10,0%
Totales	**200**	**100%**



2)

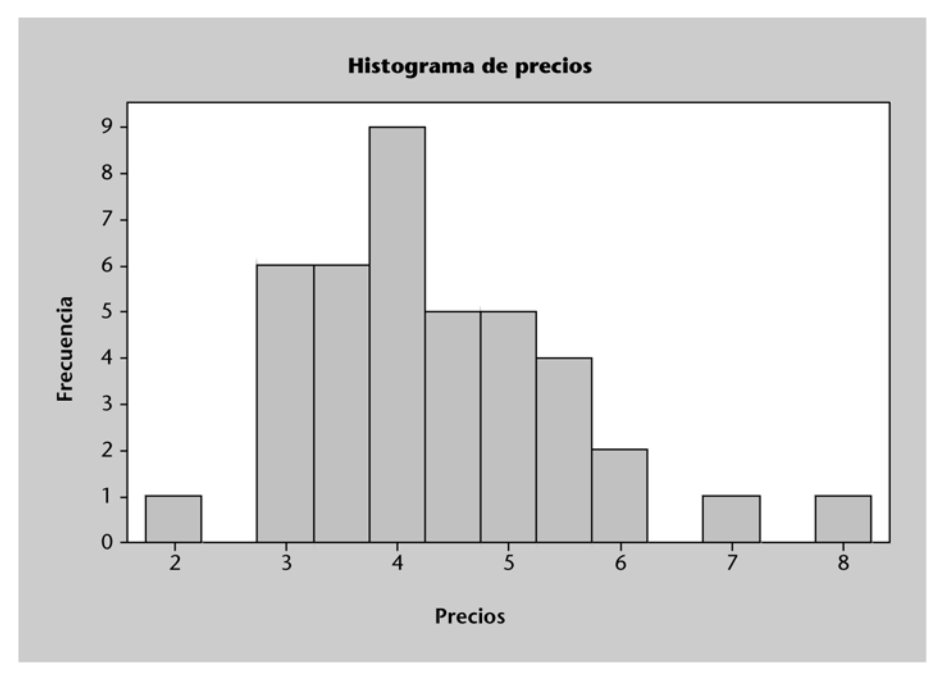
a)

```
Stem-and-Leaf Display: precios

Stem-and-leaf of precios  N  = 40
Leaf Unit = 0.10


 5    2  08899
 18   3  1155667788888
(11)  4  00013556689
 11   5  0024467
 4    6  018
 1    7  8
```

b)



c) Aunque no parece haber ningún patrón claro en los datos, sí se aprecia –tanto en el histograma como en el gráfico de tallos y hojas– una cierta forma de campana, con la parte central más elevada y unos extremos o colas más bajas. La moda de este conjunto de datos es 3,8 ya que, como se aprecia en el diagrama de tallos y hojas, es el valor que más aparece.

3)

a)

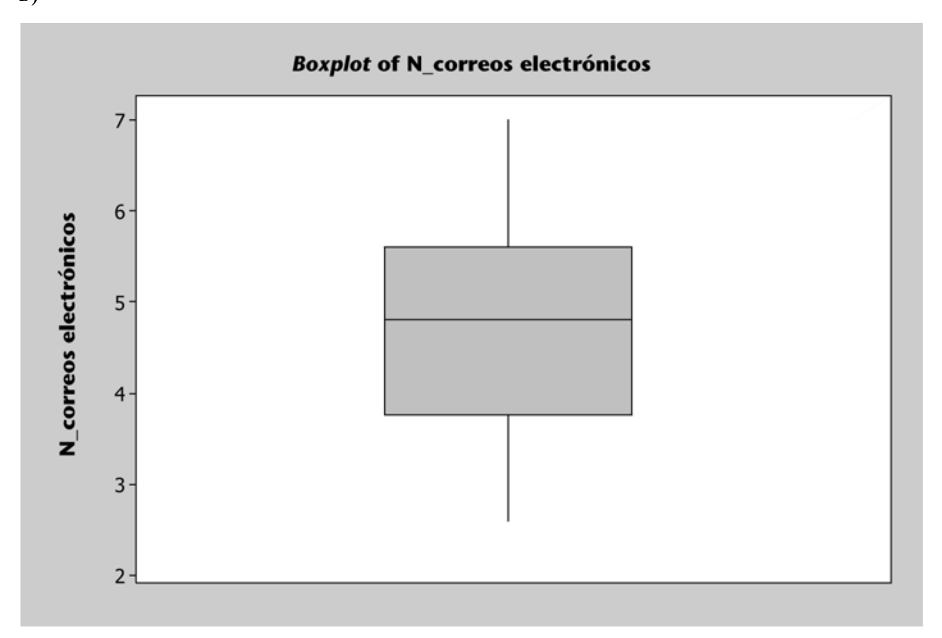
```
Descriptive Statistics: N_e-mails

Variable    N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3
N_e-mails  20   0  4.810    0.291  1.302    2.600  3.750   4.800  5.600

Variable   Maximum
N_e-mails    7.000
```

El rango intercuartílico es Q3 – Q1 = 5,60 – 3,75 = 1,85. Entre Q1 = 3,75 y Q3 = 5,60 están comprendidos el 50% de los datos centrales.

b)



No se observa, en este caso, ningún valor anómalo (*outlier*), ya que el gráfico no muestra ningún símbolo "*".

4) En este caso, puesto que el resultado de cada test puede ser "positivo" (con probabilidad 0,006) o "no positivo" (con probabilidad 1 – 0,006 = 0,994), la variable aleatoria X = "número de falsos positivos en 1.000 pruebas a atletas limpios" sigue una distribución binomial de parámetros n = 1.000 y p = 0,006. En el caso de la binomial, la media o valor esperado es $\mu = n \cdot p = 6$, es decir, cabe esperar que al aplicar el test a 1.000 atletas "limpios" haya seis falsos positivos.

Por otra parte, $P(X > 15) = 1 - P(X <= 15)$ = {Minitab o Excel} = 1 – 0,9995 = 0,0005. Por tanto, si aparecen más de quince positivos cabría pensar que muy probablemente no todos ellos sean falsos.

5) En este caso, la variable aleatoria X = "número de extranjeros en la muestra" sigue una distribución binomial de parámetros n = 1.200 y p = 0,0996. Por tanto, el valor esperado de extranjeros en la muestra es $\mu = n \cdot p$ = 119,52, es decir el promedio de extranjeros para las muestras de esas características es de, aproximadamente, 120.

Por otro lado, $P(X < 100) = P(X <= 99)$ = {Minitab o Excel} = 0,0245, es decir, es muy poco probable que una muestra contenga menos de 100 extranjeros si ésta es realmente aleatoria.

6) Se considera la variable aleatoria X = "días que dura un embarazo". Cabe tener en cuenta que $X \sim N$ (266,16).

$P(X < 240)$ = {Minitab o Excel} = 0,0521, es decir, el 5,2% de los embarazos duran menos de ocho meses.

$P(240 < X < 270) = P(X < 270) - P(X < 240)$ = {Minitab o Excel} = 0,5987 – 0,0521 = 0,5466, es decir, el 55% de los embarazos duran entre ocho y nueve meses.

Finalmente, se pide el valor c tal que $P(X > c)$ = 0,20, es decir: $P(X < c) = 1 - P(X > c)$ = 0,80 → c = {Minitab o Excel} = 279,47, es decir, el 20% de los embarazos supera los doscientos setenta y nueve días.